Universiteit Leiden
The Netherlands

Opleiding Informatica

Functional profiling of Alzheimer's through a consensus network

Noria Yousufi

Supervisor:
Dr. K.J. Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

03/07/2023

Abstract

Alzheimer's is a neurodegenerative disease, causing memory and cognitive skill problems. There is currently no effective drug or treatment and the assumption is that the number of patients with AD will only increase. Therefore research in AD is of great importance. The main objective of this research is to identify the molecular functions and biological processes which are differentially expressed in AD by means of a consensus network. Datasets from STRING, WikiPathway and KEGG have been combined into a consensus protein-protein interaction network (PPIN). Overlaying these networks with gene expression data allows for identification of differentially expressed proteins present in these PPINs. From these data the essential proteins APP, APOE, Beta-amyloid and Tau are not statistically significant and thus not differentially expressed, which was not as expected. After further research, the top three differentially expressed proteins are MPO, SELE and ERG1. With further analysis and functional enrichment analysis it can be concluded that the molecular function protein binding and the biological process response to organic substance contain the most differentially expressed proteins. Further research such as proteomics, GWAS and TWAS will allow to validate these gene expression results.

Keywords: Alzheimer's disease, protein-protein interaction networks, gene expression, functional enrichment analysis.

Contents

1 Introduction

1.1 Alzheimer's disease

Alzheimer's disease (AD) is an irreversible neurodegenerative progressive disease causing memory loss, cognitive skill problems and behavioral changes. AD is one of the well-known forms of dementia [BYBT10]. AD consists of two subvariants. Early-onset AD (EOAD), diagnosed before the age of 65, and late-onset AD (LOAD), diagnosed after the age of 65. EOAD makes up 6% of the total AD patients [ZTW+15]. There is a significant difference between these two variants [Men12]. It is therefore important to point out that this research is focused on LOAD.

There are several important proteins involved in causing LOAD, namely Amyloid Precursor Protein (APP) and Tau.
The APP is classified as a type I transmembrane protein. APP's function is to regulate the formation of synapses, which influence the communication between neurons [PBM+06]. The APPs are eventually broken down. This is done via α- and γ-secretase, which are enzymes that cut up the APPs. The formed peptide is soluble. However, in LOAD, β-secretase breaks down APP, instead of α-secretase. This results in insoluble peptides, Amyloid Beta [TAK+19]. These will build up to Beta-amyloid plaques. The Beta-amyloid plaques disrupt the signalling between neurons [LXL+21].
The protein Tau is located in the microtubule of the neuron cell. Phosphate groups bind to Tau, which changes its structure and are then released from the microtubule. This results in accumulation of Tau plaques inside the neuron, which are called neurofibrillary tangles. These can eventually provoke apoptosis of the neuron cell [Bri98].
The gene APOE $\epsilon 4$ helps to break down the Beta-amyloid protein. If this gene is disrupted it will increase the risk of LOAD. Since the Beta-amyloid proteins can not be cleaned up, these start to accumulate into plaques in the brain [KBH09].

The number of patients with dementia will increase. It is estimated that there will be 74.7 million dementia patients in 2030 and already 131.5 million in 2050. 50 to 70% of dementia is caused by AD [PWG+15]. It is therefore crucial to further research AD in order to produce an effective medicine or treatment.

1.2 Protein-protein interaction networks

In protein-protein interactions (PPIs) two or more proteins are in physical contact. Protein-protein interaction networks (PPINs) visualise the communications and interactions between these proteins, which can reveal functions and pathways within these PPINs [FSAL21]. Two interacting proteins are represented as adjacent nodes and connected via edges. A protein is considered a hub protein if its connectivity to other proteins is above the average connectivity in the network. These highly connected proteins tend to be located at the center of these networks [MLX+22]. The removal or change in gene expression levels of these hub proteins has a major impact on the network's topology, influencing the function of the proteins around them, indicating the significance of these proteins [HZ06]. PPIs and PPINs can help to identify potential drug targets and subnetworks that are associated with a specific disease. This provides understanding into the underlying mechanism of the disease [PP10] [SA14].

1.3 Research statement

The objective of this thesis project is to perform functional profiling of the protein interactions present in the PPINs, allowing to discover their activity levels and functional characteristics through network topology aiming to identify up- and downregulated protein pathways in AD through a consensus network. This thesis project aims to answer the following research question:

RQ: Which molecular functions and biological processes can be identified as differentially expressed in AD through a consensus network?

These questions will be answered through consensus PPINs, after which the hub proteins and their interactions will be identified. Since most of the important dysregulated proteins are already known, see section 1.1, particular attention is paid to how these proteins influence their first neighbours and second neighbour proteins. The results are combined with gene expression data, allowing the identification of the dysregulated pathways. Although much research has been done on dysregulated proteins in AD, it is very important to identify how these proteins influence molecular functions and biological processes in AD, which can be derived from the consensus network. The answers to these questions can help gain inside into the underlying mechanisms of AD, as new important pathways and connections may be discovered. This might contribute to potential novel drug targets.

1.4 Related work

Previous research has been conducted in which consensus PPINs were constructed for AD.
In the study of V. Srinivasa Rao et al. [RSKS13], a consensus PPIN was made. However, this was done via text mining, which is a technique that extracts information from structureless text and can be prone to errors [ZPZ+13]. This technique won't be used in this thesis project, instead direct datasets from PPIs databases were used. In addition, the study of V. Srinivasa Rao et al. hasn't identified the functional enrichment terms of molecular functions and biological processes present in the PPIN.
A more recent study from Xuemei Quan et al. [QLC+20], suggested a method to identify differentially expressed genes in AD by means of combining mRNA expression data and a constructed PPIN. However, their approach differs from this thesis project, since they used the screened genes from previous identification of differentially expressed genes that overlap with the PPIN. Hence, they do not create a consensus network from different databases, but use expression data.
Gene expression data will be applied in this thesis. In the studies of Raffaella Nativio et al. a meta-analysis was done where gene expression, proteomics and epigenomic data were obtained. They concluded disruption of chromatin regulation in AD. The two gene expression datasets of Raffaella Nativio et al. were used in this thesis: GSE104704 [NDB+18] and GSE159699 [NLD+20].

1.5 Thesis overview

The aforementioned chapter contains the Introduction Section 1. The important definitions needed to understand this thesis are provided in Section 2. Next in Section 3 the exact method followed in this thesis is described. The obtained results are displayed in Section 4. The thesis concludes with a conclusion and discussion in Section 5. Finally, further research is described in Section 7.

This bachelor thesis was conducted at LIACS (Leiden Institute of Advanced Computer Science) under the guidance of my supervisor Dr. K.J. Wolstencroft. The consensus PPINs were created in collaboration with fellow student, Aster de Boer.

2 Definitions

The following key definitions provide a clear understanding of the different terminologies and concepts used in this thesis.

- Consensus network: combines PPIs from several databases in order to come to a consensus, resulting in a more reliable and robust PPIN.
- Cut-off score: determines the number of interactions that are allowed. The lower the cut-off score the more interactions and thus more edges are visualised in the consensus network. A cut-off score of 0.4 and 0.7 was used in this thesis project. These are annotated by either consensus(0.4) or consensus(0.7).
- Network expansion: both consensus PPINs were extended further with second neighbour proteins allowing for the possibility to gain new insight into the molecular mechanism behind AD. The difference between a PPIN with expansion and without expansion is distinguished by the following notation: consensus(0.4)+expansion or consensus(0.7)+expansion.
- GEO identifier: two different datasets, GSE159699 or GSE104704, were used to overlay the consensus networks with gene expression data. The specific dataset is indicated prior to the type of consensus network. For example, GSE159699+consensus(0.4)+expansion.

3 Method

3.1 Data retrieval

Data creating the consensus network were retrieved from the databases Kyoto Encyclopedia of Genes and Genomes (KEGG), WikiPathways and STRING. The datasets were found by searching in the described databases with the keywords: "Alzheimer disease", "Alzheimer's disease", "Alzheimer's disease", "Homo sapiens", "Beta-Amyloid" and/or "Tau". Cytoscape (version: 3.9.1) was used as application tool in this thesis.

- KEGG: the dataset with entry map05010, which can be found on the KEGG website, was downloaded as KGML file. The packages `CyKEGGParser` (version 1.2.9) and `KEGGscape` (version 0.9.1) were applied.
- WikiPathways: the pathway with WP5124 was found in the WikiPathways toolbar in Cytoscape or on the WikiPathway website [Han21]. See Figure 4. Installation of the package `Wikipathways` (version 3.3.10) is needed.

- STRING: the dataset with DOID 10652 was used as STRING network. This network was retrieved by searching for 'Alzheimer" in the STRING: disease query toolbar in Cytoscape. The following settings were applied in order to import the STRING networks. The package `stringApp` (version 2.0.1) was used.
  - Select: Alzheimer's disease: DOID:10652
  - Network type: full STRING network
  - Confidence (score) cutoff: 0.4 or 0.7
  - Maximum number of proteins: 2000
  - Options: Load Enrichment Data turned on

3.2 Data preparation

In order to merge all the datasets and form a consensus PPIN all the data were preprocessed. First, the UniProt identifiers (ID) were added to the data present in the networks. The specific method of how this was done differs per dataset.

The KEGG network uses a different naming convention, namely hsa (Homo sapiens). A hsa identifier is assigned to a single protein or complexes of multiple proteins. The hsa identifiers forming a complex of multiple proteins must be split into individual proteins, after which all hsa identifiers were converted to the UniProt IDs. This process was performed by a fellow student, Aster de Boer, all necessary files for this process can be found on De Boer's GitLab: https://git.liacs.nl/s2955199/alzheimerconsensus. See Figure 5 for the visualisation of the network.

For the WikiPathway network: a `Python` (version 3.10.8) script was written to convert the column "Type" with the values "GeneProduct" to UniProt IDs, see my GitLab repository for the written code: https://git.liacs.nl/s2839334/alzheimer-WikiPathway-convert-ids. The libraries `xml.etree.ElementTree`(version 1.1.0), `unipressed` (version 1.1.0), `pyWikiPathways` (version 0.0.2) and `pandas` (version 1.5.3) were applied. This method converts 80 more Gene Products than the manual UniProt ID mapper webtool and thus incorporates more proteins in the final consensus network [uni]. The generated Excel file was imported into Cytoscape on the WikiPathway network, via File → Table from file → import columns from table (settings) → OK. The following settings were used:

- Where to import Table Data: to a network collection
- Network Collection: WikiPathway network
- Import Data as: Node Table Columns
- Key Column for network: shared name
- Case Sensitive Key Values: turned on

Second, the KEGG and WikiPathway networks were extended with the information from STRINGify, Apps → STRING → STRINGify network. This ensures that nodes and edges present in the pathway

PPINs are recognised by the STRING database. If these proteins are recognised by the STRING database then additional information is added to the pathway PPIN. The additional information contains evidence of the protein presence in different "tissues" or a different naming convention, such as "canonical name". These are relevant attributes later in this process. The settings used for creating the STRINGify networks are:

- Column for STRING query: UniProt
- Include unmappable nodes: turned off
- Map nodes to compounds: turned off
- Species for the query: Homo sapiens

As explained in the definitions section 2, the cut-off score determines the number of interactions that are allowed in the network. The lower the cut-off score the more interactions are permitted. Since a consensus network with a confidence cut-off score of 0.4 and 0.7 was created, the networks of KEGG and WikiPathway must also comply with this. This was achieved by adjusting the confidence cut-off score, by means of: Apps → STRING → Change confidence or type. The settings used were:

- Confidence cut-off score: optionally 0.4 or 0.7
- Network type: full STRING network

Ultimately, the empty values in the "canonical name" column must be removed in all networks, including the STRING(0.4) and STRING(0.7) networks. If this is not removed, Nan values will arise which can cause problems with merging the networks. These values were removed by means of a filter, Filter → click plus sign → choose column filter → select Node: canonical name → select doesn't contain → Apply. The nodes without a canonical name were deleted by: Edit → Remove selected nodes and edges.

Figure 6a visualises the KEGG(0.4) and Figure 6b presents the KEGG(0.7) networks. Reference is made to Figure 7a for the WikiPathway(0.4) and Figure 7b for the WikiPathway(0.7) networks. Finally, Figure 8a, presents the STRING(0.4) and Figure 8b the STRING(0.7) networks. See Figure 1 for a workflow of the followed method.

3.3 Creating consensus network

The constructed consensus PPINs are undirected graphs, where nodes represents proteins and edges represent the interactions between the proteins. PPINs are considered scale-free, meaning some proteins have a high degree of connectivity which are considered hub proteins, but most proteins in the PPIN have a low degree of connectivity. The cut-off score represents the amount of evidence needed in order to provide an interaction. The lower the cut-off score the more weaker interactions are allowed.

The different networks were merged to create consensus(0.4) and consensus(0.7) PPINs. This was done in Cytoscape using: Tools → Merge → Networks. Union ensures that all information from all data sets is merged. The "canonical name" column was used to merge the networks together, at

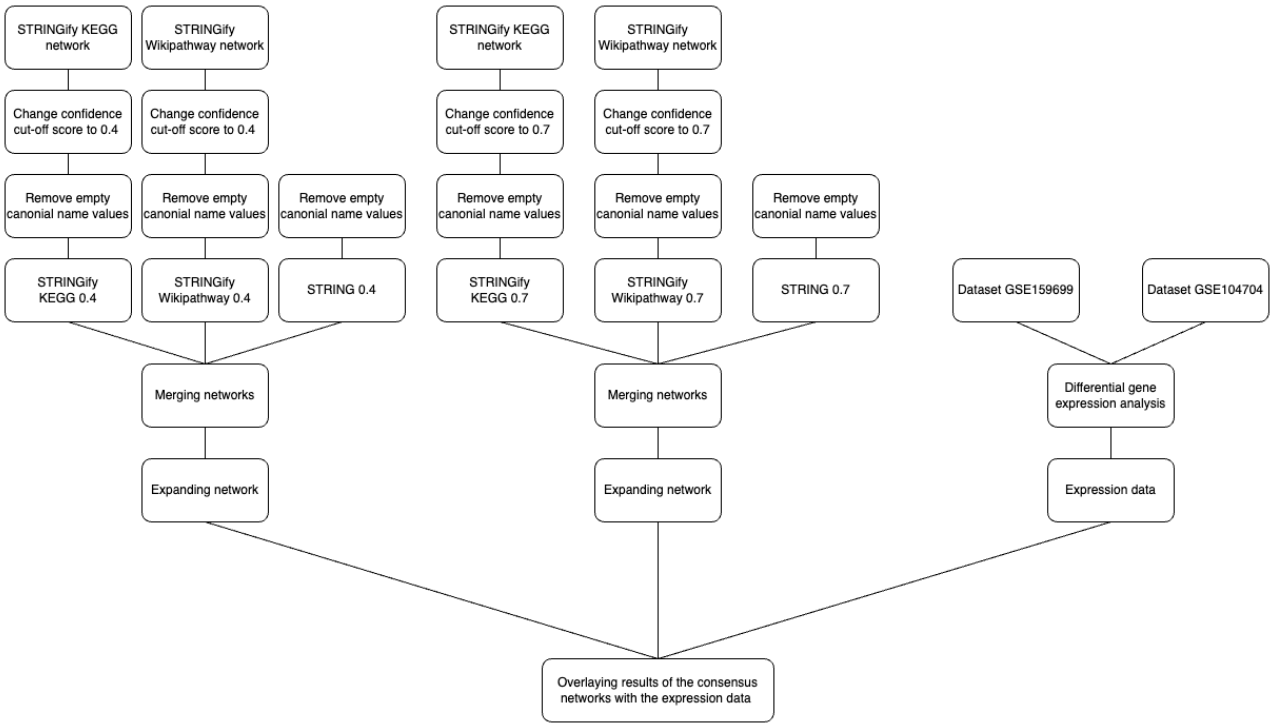


Figure 1: Workflow of the followed method.

the Advanced Option tab.

Subsequently, additional proteins were added to the consensus network. This further expansion allows for more captured interactions and may lead to new insight into AD, since these PPI aren't captured by the datasets used in the consensus networks. Both networks were expanded by: Apps → STRING → Expand network. The following settings were used:

- Number of interactors to expand network by: 2000
- Type of interactors to expand network by: Homo sapiens
- Selectivity of interactors: 0.5

This added two thousand second neighbour proteins to the network. Second neighbours share interactions indirectly with the proteins already present in the network. These proteins expand the original consensus network and thus may reveal new important protein pathways and drug targets.

Both networks must be filtered for the column: "tissue nervous system", providing evidence for the presence of the protein in the nervous system. This attribute was chosen since it is accepted that most proteins involved with AD are located in the nervous system. However, the fact that there is no evidence doesn't imply that the proteins aren't part of the nervous system. Since a few proteins didn't contain any evidence, it was decided that those proteins weren't removed.
See Figure 10a for the consensus(0.4) network and see Figure 10b for the consensus(0.7) network. Figure 9 visualises the consensus PPINs without STRING expansion.

3.4 Network Analysis

The topology of the networks was analysed with the Analyze tool in Cytoscape. See: Tools → NetworkAnalyzer → Network Analysis → Analyze. Choose as Network Interpretation: Treat the network as undirected. Using this method, the hub proteins in the networks can be identified based on the "degree" indicating the connectivity of the protein. The analysed topology features are:

- Number of nodes.
- Number of edges.
- Average number of neighbours.
- Network diameter: the longest distance between any pair of two nodes.
- Network radius: the shortest path between the most remote nodes of the PPIN.
- Characteristic path length: is the average length of the shortest route between each pair of nodes.
- Clustering coefficient: average of the number of connections the neighbours of a node have with each other in the whole PPIN.
- Network density: the extent to which a network is fully connected by comparing the proportion of the number of interactions a node can have with the maximum number of possible interactions the node could have.
- Network heterogeneity: the likelihood that a network incorporates hub proteins.
- Network centralization: the measure of nodes that have a much higher degree than other nodes in the network.
- Connected components: subgraphs in the network where all the nodes are fully connected [Jia22].

See Table 1 and 2 for the results.

3.5 Expression data

By using expression data, it is possible to identify up- or downregulated proteins present in AD. These proteins are dysregulated in AD compared to a healthy control group. This is achieved by overlaying these results with the consensus networks, which allows to identify differentially expressed proteins.
The data used for this process were obtained from *GEO RNA-seq Experiments Interactive Navigator* (GREIN). GREIN is an interface where GEO RNA-seq data can be analysed [MNP+19]. These datasets are originally published on *Gene Expression Omnibus* (GEO), which is a database for gene expression data [EDL02]. The data was found by searching for "Alzheimer's" and "Alzheimer's disease" in the search bar on GREIN's website. The dataset with GEO accession numbers: GSE159699 and GSE104704 was applied. These datasets were selected based on the comparison made between

healthy elderly people and patients with AD, where no other conditions are implied. In addition, these datasets were recently released.

The GSE159699 dataset contains 30 samples of elderly people, young people and AD patients. These samples were taken from the hippocampus [NLD+20]. The GSE104704 dataset also contains 30 samples comparing the elderly with AD patients. These samples were obtained from the lateral temporal lobe. In both dataset RNA-seq was applied as a high throughput method [NDB+18].

The datasets used were obtained by: Analyze dataset → Factor of interest: characteristics → Sample selection: Specific samples → Select experimental samples: all the AD samples → Select control samples: all the old samples → Type of comparison: Two group without covariate → Generate signature.
The consensus networks contain an *ensemble protein id* (ENSP). The generated data set from GREIN consists of an *ensemble gene id* (ENSG). A `Python` (version 3.10.8) script was made to convert the ENSP values in the consensus networks to ENSG values. The libraries used were `pandas` (version: 1.5.3) and `Biomart` (version: 0.9.2). The code is available on my GitLab: https://git.liacs.nl/s2839334/alzheimer-WikiPathway-convert-ids. These ENSG values were then added to the consensus networks, after which the GREIN dataset could be merged on the ENSG value.
There are several values in the dataset present. The *logarithm of fold change* (logFC) is an indication of the variation in gene expression levels under different conditions [DK14]. A positive value indicates upregulation of the gene and negative values indicate downregulation of the gene. By performing multiple testing, which is the case with measuring gene expression levels, the likelihood of false positive results increases. The *adjusted p-value* corrects the original p-value for multiple testing, which decreases the rate of false positives. A gene is considered statistically significant and differentially expressed if the FDR adjusted p-vale is below the threshold of 0.05. [PMK+05].

3.6 Clustering

Clustering the networks identifies highly connected subnetworks. The clustering of networks also identifies different biological processes and functions through the connection with other proteins within the subnetwork [PR14]. The MCODE extension in Cytoscape enables to cluster networks. `MCODE` version 2.0.2 was applied using the default settings.

3.7 Functional enrichment analysis

The derived clusters were functionally enriched with the annotation database *Gene Ontology* (GO). The proteins can be categorized into the following components: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). In this research, the MFs and BPs are further investigated. This allows to investigate which MFs and BPs are overrepresented in the clusters. Only the clusters containing one or more of the important proteins, see section 1.1, were further investigated and used for functional enrichment analysis.
This was achieved by making the cluster in question a STRING network by: Apps → STRING → STRING Network. Subsequently, the functional enrichment analysis could be performed by: STRING enrichment → Retrieve functional enrichment. Thereafter only the Molecular function or Biological process was chosen in the Filter STRING Enrichment table.

The *False discovery rate* (FDR) identifies which enriched GO terms are significantly present in the cluster. The FDR limits the number of false positives and therefore increases the confidence of the results. The lower the FDR value the more significant the result. The top five MFs and BPs with the lowest FDR values are visualised in the clusters.

4 Results

4.1 Consensus network

The expanded consensus 0.4 and 0.7 networks are shown respectively in Figure 10a and 10b. The important proteins such as APP, Beta-Amyloid, APOE and Tau are coloured red. The other colours belong to the respective dataset or overlapping datasets. The node size depends on the "tissue nervous system" column. This value indicates the probability that the protein is located in the nervous system. The more evidence the larger the node size of the protein and vice versa.

4.2 Network Analysis

The number of nodes per database is shown in Table 1, these numbers are the same for the expanded consensus(0.4) network as for the expanded consensus(0.7) network, since the confidence cut-off score only affects the number of interactions that are allowed, thus the number of edges in the network and not the number of proteins.

Table 1: Results of the number of nodes per database in the expanded consensus PPINs.

Database	**Number of nodes**
KEGG	84
STRING	1726
WikiPathway	1
STRING;WikiPathway	1
STRING;KEGG	41
WikiPathway;KEGG	115
STRING;WikiPathway;KEGG	136
STRING expansion	2000

Table 2 shows the network analysis of the different topology features that Cytoscape measures. Both networks have the same number of proteins and only differ in the number of edges. The differences in the topology features can therefore only be explained by this property. Based on Table 2 it is apparent that a higher confidence cut-off score results in a lower average number of neighbours, network density and network centralisation. On the contrary, a higher confidence cut-off score results in higher values for the other topology features.

Table 2: Results of the network topology analysis on both consensus expanded PPINS.

	Consensus(0.4)+expansion	**Consensus(0.7)+expansion**
Number of nodes	4104	4104
Number of edges	256599	96834
Avg. number of neighbours	124.564	47.094
Network diameter	5	8
Network radius	3	4
Characteristic path length	2.286	2.903
Clustering coefficient	0.356	0.416
Network density	0.030	0.012
Network heterogeneity	0.885	0.962
Network centralisation	0.312	0.112
Connected components	1	33

4.3 Gene expression data

4.3.1 Visualisation and analysis

The two datasets, GSE159699 and GSE104704, are overlayed with the consensus(0.4) and consensus(0.7) PPINs. As shown in Table 3, the GSE159699 dataset contains 610 genes with an FDR adjusted p-value below 0.05. These genes are significantly differentially expressed in AD patients compared to elderly people who do not have AD. Overlaying this dataset on the PPINs, identified 154 proteins in the consensus(0.4)+expansion and 153 proteins in the consensus(0.7)+expansion PPINs that are significant. Similarly, the GSE104704 dataset contains 702 genes with an FDR adjusted p-value below 0.05, out of which 174 proteins are in the consensus(0.4)+expansion PPIN. The consensus(0.7)+expansion PPIN contains 173 significant proteins. Therefore, there is a slight difference in the number of significant proteins present in different datasets and thus also in the PPINs.

Table 3: Number of significant genes present in the different datasets and PPINs.

	Number of Genes (FDR p <0.05)	**Consensus(0.4)+expansion PPIN**	**Consensus(0.7)+expansion PPIN**
GSE159699	610	154	153
GSE104704	702	174	173

As shown in Table 3 the cut-off score does not influence the amount of significant proteins present in the PPIN. Since the cut-off score only influences the number of interactions and not the number of proteins present. However, a difference of one protein is present, probably due to the STRING expansion done on the PPINS. This adds different second neighbour proteins and thus a small difference in the presence of the proteins in the consensus expanded networks.

The expression levels of the datasets are visualised in heatmaps, which provide an understanding of the gene expression patterns among different conditions. The heatmaps clearly visualise the different clusters of up- and downregulated genes indicating a difference in expression patterns of

AD in comparison with healthy elderly people. See Figure 11 for the heatmap of the GSE159699 dataset and Figure 12 for the GSE104704 dataset.

The overlayed gene expression datasets are represented in Figure 2 and Figure 3 where the colour gradient corresponds to the LogFC value, indicating the up- or downregulation of the protein. Downregulation is represented in green and upregulation in red. The large nodes contain a FDR adjusted p-value below 0.05 and are therefore considered statistically significant. These proteins are potentially interesting for this research since they indicate a strong probability of involvement in AD.
The important proteins associated with AD, namely APP, Beta-amyloid, Tau and APOE, were identified in Section 1.1. They are represented as large triangles in the visualisations to stand out. Unexpectedly this analysis reveals that these proteins don't have a FDR adjusted p-value below 0.05. Therefore these proteins don't provide enough evidence for their dysregulation in AD and thus are not statistically significant. This is surprising since these proteins are considered as the main cause behind AD in literature research.
A comparison is made to analyse the presence of up- and downregulated proteins between the two datasets, as shown in Table 4. There are overall more downregulated proteins present in both datasets and thus also in the different PPINs. There are more statistically significant downregulated proteins present, than upregulated proteins in all PPINs. Logically the GSE104704 dataset maps more significant proteins on the PPINs since this dataset contains more statistically significant genes.

Table 4: Representation of the number of up- or downregulated proteins present in each PPIN. In addition, the number of statistically significant proteins present in each PPIN are represented.

Dataset	**Consensus Network**	**Upregulated Genes**	**Downregulated Genes**	**FDR p<0.05 and Upregulated Genes**	**FDR p<0.05 and Downregulated Genes**
GSE159699	Consensus (0.4) + expansion	1715	2170	39	114
GSE104704	Consensus (0.4) + expansion	1528	2270	33	141
GSE159699	Consensus (0.7) + expansion	1690	2179	42	112
GSE104704	Consensus (0.7) +expansion	1527	2345	36	137

The hub proteins were identified in both consensus PPINs. A hub protein has a higher connectivity than the average connectivity present in the PPIN. The top five hub proteins present in both PPINs are mentioned in Table 5 and Table 6. A higher degree indicates higher connectivity within the network, therefore these tables are ordered in descending order of the degree value. The FDR adjusted p-value indicates that these top five hub proteins are not considered differentially expressed. In addition, the LogFC value indicates whether these are up- or downregulated. The essential proteins APP, APOE, Beta-amyloid and Tau are not part of the top five most connective proteins

present in the PPIN. The hub proteins are highly connective in the PPIN, but are not considered differentially expressed and are therefore not further considered in this research.
The top five most statistically significant differentially expressed genes in the GSE159699+consensus (0.4)+expansion PPIN are represented in Table 7 and the GSE104704+consensus(0.4)+expansion PPIN in Table 8. None of the top five differentially expressed proteins are considered hub proteins, since their degree does not exceed the average number of neighbours in the consensus(0.4)+expansion of 124.564. Reference is made to Table 2.
The top five differentially expressed genes of the consensus(0.7)+expansion overlayed with the GSE159699 and GSE104704 dataset are respectively represented in Table 9 and Table 10. Likewise, none of these top five genes corresponding to a protein are considered hub proteins, since their degree does not exceed the average number of neighbours present in the network of 47.095. Reference is made to Table 2.

Table 5: Top five hub proteins present in the consensus(0.4)+expansion PPIN. The table is ordered in descending order of the degree.

Hubs	**Degree**	**Adjusted p-value GSE159699**	**Adjusted p-value GSE104704**	**LogFC GSE159699**	**LogFC GSE104704**
AKT1	1408.0	0.835	0.925	0.079	0.046
ACTB	1316.0	-	-	-	-
GAPDH	1287.0	0.123	0.093	-0.439	-0.473
TP53	1175.0	0.780	0.712	-0.12	-0.149
TNF	1133.0	-	0.820	-	-0.785

Table 6: Top five hub proteins present in the consensus(0.7)+expansion PPIN. The table is ordered in descending order of the degree.

Hubs	**Degree**	**Adjusted p-value GSE159699**	**Adjusted p-value GSE104704**	**LogFC GSE159699**	**LogFC GSE104704**
TP53	504.0	0.780	0.712	-0.12	-0.149
SRC	493.0	0.884	0.831	-0.091	-0.126
RPS27A	456.0	0.617	0.681	0.228	0.196
CTNNB1	453.0	0.693	0.605	-0.131	-0.165
AKT1	415.0	0.835	0.925	0.079	0.046

Table 7: Top five differentially expressed genes present in the GSE159699+consensus(0.4)+expansion PPIN ascending order of the FDR adjusted p-value.

Differentially expressed gene	Degree	Adjusted p-value GSE159699	LogFC GSE159699
VGF	41.0	7.98E-13	-2.428
RPH3A	44.0	4.68E-10	-1.502
CRH	121.0	1.04E-07	-2.447
PCSK1	67.0	9.08E-06	-1.891
NEUROD6	75.0	9.08E-06	-1.732

Table 8: Top five differentially expressed genes present in the GSE104704+consensus(0.4)+expansion PPIN ascending order of the FDR adjusted p-value.

Differentially expressed gene	Degree	Adjusted p-value GSE104704	LogFC GSE104704
VGF	41.0	1.15E-12	-2.465
RPH3A	44.0	2.56E-10	-1.535
CRH	121.0	6.99E-08	-2.48
DUSP6	71.0	3.88E-06	-0.964
PCSK1	67.0	5.89E-06	-1.929

Table 9: Top five differentially expressed genes present in the GSE159699+consensus(0.7)+expansion PPIN ascending order of the FDR adjusted p-value.

Differentially expressed gene	Degree	Adjusted p-value GSE159699	LogFC GSE159699
VGF	10.0	7.98E-13	-2.428
CRH	34.0	1.04E-07	-2.447
NEUROD6	1.0	9.08E-06	-1.732
PCSK1	9.0	9.08E-06	-1.891
MPO	32.0	3.20E-05	-2.098

4.4 Clustering & Functional enrichment analysis

Clusters were obtained from the consensus+expansion PPINs. The expanded consensus(0.4) network yielded 31 clusters and the expanded consensus(0.7) resulted in 62 clusters. In total four clusters were obtained containing one or more important proteins APP, APOE, Beta-amyloid and Tau, identified in section 1.1. These clusters were functionally enriched with their MFs and BPs. The MFs and BPs are represented by means of a donut chart around the protein. These are sometimes filled with one or more colours corresponding to a function or process. Sometimes proteins do not contain colours, because the function or process they belong to does not belong to the top five most

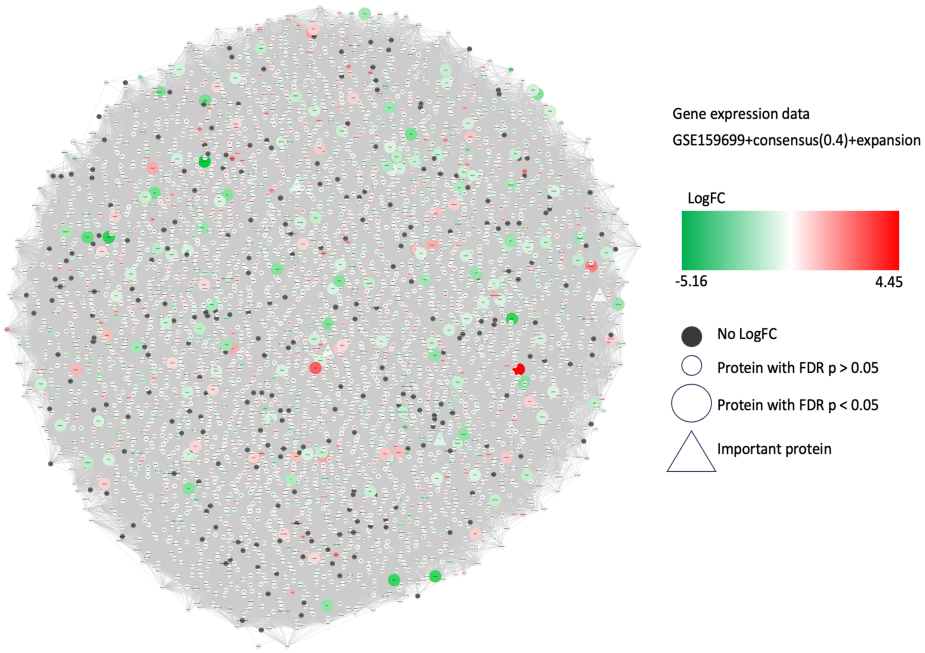
Table 10: Top five differentially expressed genes present in the GSE104704+consensus(0.7) +expansion PPIN on ascending order of the FDR adjusted p-value.

Differentially expressed gene	**Degree**	**Adjusted p-value GSE104704**	**LogFC GSE104704**
VGF	10.0	1.15E-12	-2.465
CRH	34.0	6.99E-08	-2.48
PCSK1	9.0	5.89E-06	-1.929
NEUROD6	1.0	6.63E-06	-1.763
MPO	32.0	2.48E-05	-2.135

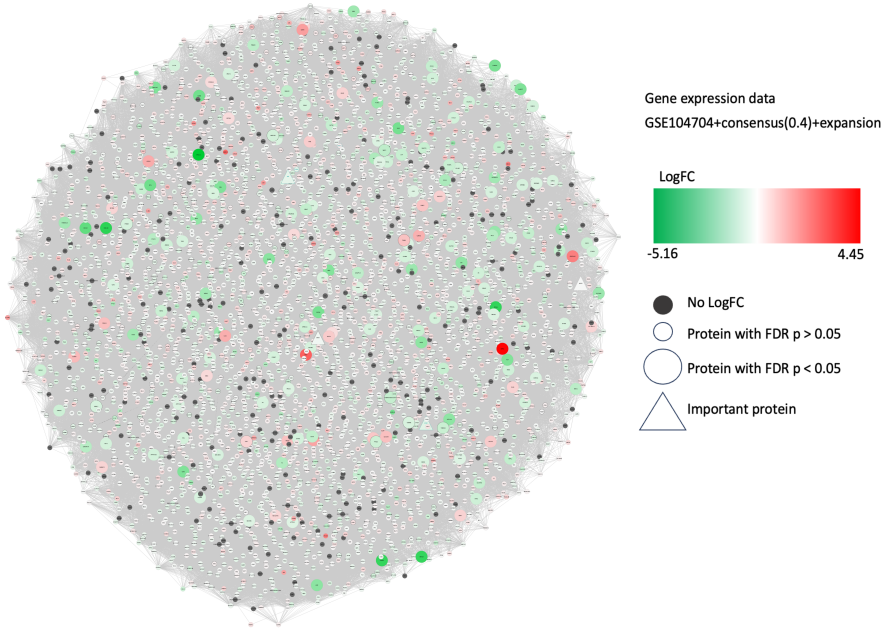
common functions or processes. Cluster 1 from the consensus(0.7)+expansion network is displayed in Figure 13. Cluster 5 from the consensus(0.7)+expansion network is visualised in Figure 14. These clusters are overlayed with gene expression data, however their images are not shown, since these clusters are not used for further analysis. The clusters from the consensus(0.4)+expansion are used for further analysis and are therefore visualised with their overlayed gene expression data. The motivation behind this will be explained later on. Cluster 2 and 3 with overlayed gene expression data from the consensus(0.4)+expansion are visualised in Figure 15 and Figure 16 respectively.

These four clusters were overlayed with the gene expression datasets GSE159699 and GSE104704. An analysis was performed on all the proteins containing a FDR adjusted p-value below 0.05 in order to identify the most statistically significant disrupted proteins in these clusters. The results of this analysis are shown in Table 11. As mentioned earlier, the important proteins are not statistically significant and therefore interest is shifted toward significantly disrupted proteins present in the clusters containing one or more of these proteins. The top 15 proteins are represented based on ascending order of the FDR adjusted p-value. The complete table containing this information on all the proteins present in identified clusters below the FDR $p<0.05$ threshold is present on the Gitlab: https://git.liacs.nl/s2839334/alzheimer-WikiPathway-convert-ids.

This shortened table version displays only proteins present in the clusters from the consensus(0.4)+expansion PPIN. Due to more proteins present in the clusters from this PPIN, the 0.4 cut-off score allows for more interactions, creating larger clusters. Since more proteins present allow for more proteins to be mapped onto the datasets and thus a higher amount of proteins containing a FDR $p<0.05$.
From this table, the three most important proteins can easily be identified. Namely, MPO and SELE in cluster 2 derived from the consensus(0.4)+expansion, their significance is supported by both datasets. In addition, EGR1 is present in cluster 3 also derived from the consensus(0.4)+expansion and considered statistically significant in both datasets. As indicated by the LogFC column these genes are all downregulated in AD. The top three differentially expressed proteins present in the clusters are further analysed.

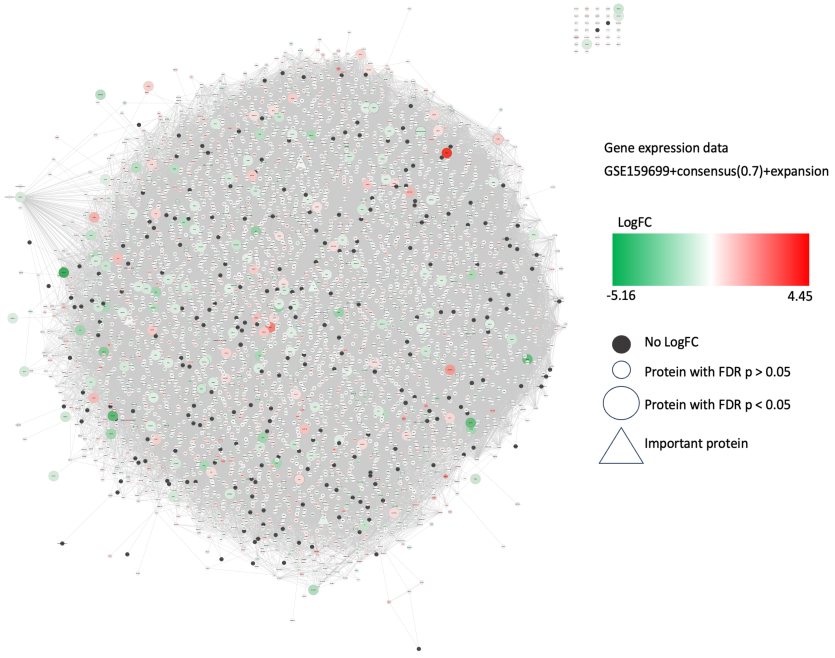


(a) Gene expression dataset GSE159699 overlayed on expanded consensus 0.4 PPIN.

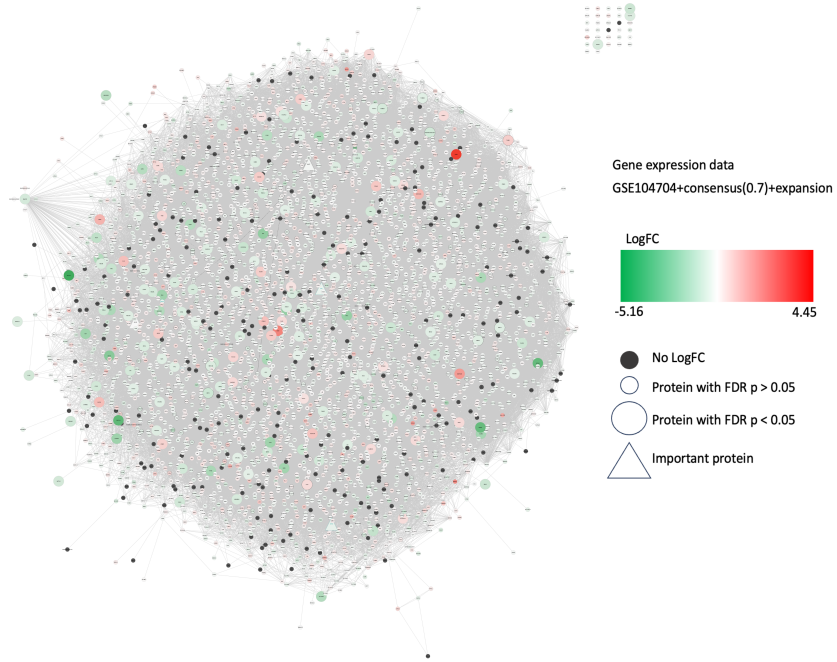


(b) Gene expression dataset GSE104704 overlayed on expanded consensus 0.4 PPIN.

Figure 2: Gene expression data overlayed on the consensus(0.4)+expansion PPIN. The LogFC value is visualised with a colour gradient, were downregulated proteins are represented in green and upregulated proteins in red. The important proteins: APP, APOE, Beta-amyloid and Tau are visualised as large triangles. Proteins without LogFC value are indicated in black.



(a) Gene expression GSE159699 overlayed on expanded consensus 0.7 PPIN.



(b) Gene expression GSE104704 overlayed on expanded consensus 0.7 PPIN.

Figure 3: Gene expression data overlayed on the consensus(0.7)+expansion PPIN. The colour gradient represents the LogFC value, where the downregulated proteins are indicated in green and upregulated proteins in red. The important proteins: APP, APOE, Beta-amyloid and Tau are visualised as large triangles.

Table 11: Representation of the top 15 most statistically significant differentially expressed proteins present in the derived clusters. Each protein is indicated with the corresponding cluster and original network type, either the consensus(0.4)+expansion indicated with 0.4 or the consensus(0.7)+expansion indicated as 0.7. The corresponding FDR adjusted p-value and LogFC value are indicated based on their derived dataset. The number of MFs and BPs, derived from the functional enrichment analysis, in which these proteins are involved is indicated. Enrichment analysis composes the top five enriched MFs or BPs. So, the number of MFs or BPs mapped to a protein can not exceed 5.

Gene Name	Dataset Presence	Cluster	Original network cut-off score	FDR Adjusted p-value	LogFC	Amount of MF pathways	Amount of BP pathways
MPO	GSE104704	2	0.4	2.48E-05	-2.135	0	1
MPO	GSE159699	2	0.4	3.20E-05	-2.098	0	1
SELE	GSE104704	2	0.4	1.28E-04	-2.708	1	3
SELE	GSE159699	2	0.4	1.56E-04	-2.681	1	3
EGR1	GSE104704	3	0.4	3.01E-04	-1.911	2	5
EGR1	GSE159699	3	0.4	3.90E-04	-1.858	2	5
HSPB3	GSE104704	3	0.4	8.06E-04	-1.739	0	1
CXCL11	GSE104704	2	0.4	8.95E-04	-5.159	5	4
HSPB3	GSE159699	3	0.4	9.00E-04	-1.706	0	1
CXCL11	GSE159699	2	0.4	1.01E-03	-5.16	5	4
CXCL10	GSE104704	2	0.4	1.05E-03	-4.275	5	5
CXCL10	GSE159699	2	0.4	1.30E-03	-4.278	5	5
CPLX1	GSE104704	3	0.4	1.33E-03	-1.048	2	0
CPLX1	GSE159699	3	0.4	1.94E-03	-1.016	2	0
IL33	GSE104704	2	0.4	2.05E-03	-1.229	5	5

4.4.1 MPO

The MPO gene encodes for the myeloperoxidase enzyme producing reactive oxygen species (ROS). ROS is damaging to the neuron cells. As shown in Table 5 functional enrichment analysis indicates that MPO is not involved in any MF. However, it is involved in one BP: response to organic substance. The disruption of MPO may influence the response to the release of organic substance, Beta-amyloid, by releasing ROS. This disruption leads to aggregation of the Beta-amyloid protein in AD, since the Beta-amyloid accumulation seems to worsen in oxidative conditions [Smi98]. MPO is considered downregulated, which is supported by both datasets with a negative LogFC value. However, research suggests that its up- or downregulation can vary in gender [RRM+99].

4.4.2 SELE

The SELE gene, also known as E-selectin, is involved in the MF protein binding. The E-selectin protein is involved in three BPs, namely regulation of immune system process, response to cytokine and response to organic substance. E-selectin is expressed by endothelial cells. This protein is involved in the immune response and inflammation caused by Beta-amyloid plaques in AD [WN12].

This response activates E-selectin, which migrates leukocytes to the brain [RMN+05]. This protein is considered to be downregulated, which is indicated by Table 11.

4.4.3 EGR1

The EGR1 gene, also known as the early growth response 1 protein, is a transcription factor that can control the gene expression of various other genes that are involved in neural plasticity and snaptic activity [HCH+19]. Functional enrichment analysis determined the involvement of the two MFs protein binding and binding. These MFs correspond to the transcription factor role of EGR1 and therefore need to bind to DNA and other proteins [Lat93]. The five BPs involved are response to organic substance, cell surface receptor signalling pathway, positive regulation of biological process, cellular response to chemical stimulus and cellular response to organic substance. These BPs indicate EGR1's involvement in neural plasticity and snaptic acitivity. Table 11 indicates the downregulation of EGR1 in AD patients, which causes degradation of neural plasticity and may affect cognitive abilities [PTL20].

4.5 Dysregulated functions and processes in clusters

To narrow down the results it has been decided to investigate the clusters from the consensus(0.4)+expansion further since the most dysregulated genes are present in the clusters derived from this network, as shown in Table 11. In addition, only the GSE159699 dataset on these clusters is examined in detail. The GSE159699 dataset specifically mentions samples taken from the hippocampus. This is the brain region most affected by AD [DA11]. In addition, this dataset is more recent, therefore GSE159699 is the preferred dataset for further investigation. This retains two clusters to analyse with associated MF and BP, which are shown in Figure 15 and Figure 16 and further explained in Section 4.5.1 and 4.5.2.

An overview has been made in Table 12 to determine the most dysregulated MF in AD. This indicates that the MF "protein binding" contains the most differentially expressed proteins and is therefore the most dysregulated MF present in these clusters. This table indicates downregulation of this MF. As shown in Table 13, representing the most disrupted BP present in these clusters, the BP "response to organic substance" contains the most significant differentially expressed downregulated proteins.

4.5.1 Cluster 2 MF and BP

The retrieved functional enrichment analysis of the MFs and BPs of cluster 2 from the GSE159699 +consensus(0.4)+expansion PPIN is visualised respectively in Figure 15a and Figure 15b. The MPO and SELE proteins are present in this cluster, they are the top two most differentially expressed genes present among the clusters, see Table 11. These proteins are downregulated and in direct contact with the APOE protein, which may suggest an influence on the APOE protein. Overall, the APOE protein is in contact with 4 differentially expressed proteins, from which three downregulated and one upregulated. The other important protein present in this cluster is Tau, represented by MAPT. This protein is in contact with one significant downregulated protein.
While the MPO protein is not mapped to any GO MF, the SELE protein is involved in protein

Table 12: Representation of the MFs containing the most differentially expressed genes overlayed with the GSE159699 dataset.

Cluster	**Molecular function**	**Up regulated Genes**	**Down regulated genes**	**FDR p<0.05 and upregulated genes**	**FDR p<0.05 and downregulated genes**	**Total FDR p<0.05**
2	Protein binding	106	128	3	12	15
3	Protein binding	175	193	4	9	13
3	Binding	203	228	4	9	13
2	Signaling receptor binding	40	52	2	5	7
2	Cytokine activity	16	28	2	4	6

Table 13: Representation of the BPs containing the most differentially expressed genes overlayed with the GSE159699 dataset.

Cluster	**Biological process**	**Up regulated Genes**	**Down regulated genes**	**FDR p<0.05 and upregulated genes**	**FDR p<0.05 and downregulated genes**	**Total FDR p<0.05**
2	Response to organic substance	84	123	3	12	15
2	Response to cytokine	56	90	3	10	13
3	Response to organic substance	119	113	4	8	12
2	Cellular response to cytokine stimulus	56	85	3	8	11
3	Positive regulation of biological process	160	163	3	8	11

binding. In addition, APOE is also involved in protein binding. The downregulation of the SELE gene may disrupt the APOE protein, which clears the Beta-amyloid plaques, as a result of which APOE may have difficulty binding to these plaques resulting in the accumulation of these plaques and worsen the progression of the disease [SWH+93].
In terms of GO BPs is the MPO protein involved in response to organic substance. The SELE protein is involved in response to cytokine, regulation of immune system process and response to organic substance. These BPs refer to a response caused when Beta-amyloid plaques are build up in the brain, which are considered organic substances, which cause an immune response by releasing cytokines [PPM+05]. To map the available MFs and BPs, a treemap of the top 15 available GO terms has been visualised using the web tool Revigo [SBŠŠ11], reference is made to Figure 17a and Figure 17b.

4.5.2 Cluster 3 MF and BP

The MFs and BPs from cluster 3 are visualised in Figure 16a and Figure 16b respectively. From Table 11 EGR1 has been identified as a highly differentially expressed gene. The only essential protein present in cluster 3 from the GSE159699 +consensus(0.4)+expansion PPIN is APP. However, EGR1 does not share a direct connection with APP. EGR1 shares a connection with BDNF, another significantly differentially expressed gene, of which its protein is directly connected with APP. Therefore EGR1 and APP are second neighbours in the cluster.
EGR1 is involved in the GO MFs protein binding and binding. In addition, BDF and APP are also part of these MFs pathways. Therefore the highly disrupted EGR1 which is downregulated may be a key contributor to the dysregulation of protein binding and binding in AD [HLL+22].
The EGR1 protein is involved in all the top five BPs present in cluster 3 after functional enrichment analysis. This also accounts for the BDNF and APP proteins. Namely, response to organic substance, cell surface receptor signaling pathway, postitive regulation of biological process, cellular response to chemical stimulus and cellular response to organic substance. These processes refer to the response of the body to the release of Beta-amyloid plaques. Resulting in disregulation of the signalling pathways in the brain causing communication problems. The top 15 MFs and BPs present in this cluster is visualised in a tree map obtained from Revigo. Reference is made to Figure 18a and Figure 18b.

5 Conclusions

This research aimed to identify the disputed molecular function (MF) and biological process (BP) pathways present in AD by means of the creation of consensus PPINs integrated with gene expression data. Thereafter importance is layed onto the essential proteins APP, Beta-Amyloid, APOE and Tau, which weren't considered differentially expressed by the two gene expression datasets GSE159699 and GSE104704. However, other statistically significant differentially expressed proteins, such as MPO, SELE and EGR1 were found within the clusters. These proteins are involved in disrupting several MFs and BPs pathways identified by functional enrichment analysis. The MF pathways identified were involved in protein binding and binding. The number of differentially expressed proteins may disrupt APOE binding to Beta-amyloid plaques, as result of which APOE's inability to clear these plaques results in more accumulation. Additionally, they contribute to the key BP pathway response to organic substances, referring to the response caused by the release of Beta-amyloid plaques in the brain. This response to inflammation releases cytokines.

6 Discussion

The main limitation of the conducted research is the presence of surprising results in the two used datasets as the essential proteins APP, APOE, Beta-amyloid and Tau in AD were not differentially expressed. In addition, some genes were downregulated where literair research indicates upregulation. There are several reasons for these possible results. Namely, the brain region from which the samples are taken can have an impact on gene expression levels [TJWJ11]. In addition, it should be taken into account that most samples were taken from male subjects. Gene expression levels in the brain can differ between genders [TRI+13]. For example, the MPO gene expression level works opposite

between women and men [RRM+99]. In addition, both datasets are gathered by the same research group. The method followed and the number of samples taken are almost identical, resulting in similar results. Therefore validation is essential and different methods are discussed in further research.

7 Further Research

This research can be expanded in various ways. Firstly, proteomics data can be applied to the already used gene expression data. Overlaying the consensus PPINs with proteomics data will provide a new understanding of functional modifications present in the PPIs. Proteomics data validate the findings of the gene expression results since the functional modifications identified by proteomics data will identify which genes were disrupted. In addition, proteomics data allows for identifying post-translational modifications (PTM). These modifications, such as phosphorylation, play a crucial role in AD [YLCC20] [CQ09]. Therefore proteomics data may give insight into these PTMs in AD.
Genome-Wide Association Studies (GWAS) can also be applied as a further extension of this research. GWAS identifies significant genetic variants associated with AD, which are responsible for the dysregulation present in the consensus PPINs [LK14]. Using GWAS data enables easy comparison between various diseases. This may give insight into the corresponding underlying mechanism of AD and other diseases, such as Huntington's or Parkinson's disease.
Transcriptome-Wide Association Study (TWAS) might provide a further understanding of the gene expression levels of the identified genetic variants discovered by the GWAS analysis [ZDJZ21]. This might result in newly identified key genes and thus new pathways in the PPINs. TWAS data may discover the reliability or strengthen the results from the gene expression data applied in this thesis.

8 Acknowledgement

I would like to express my sincere gratitude for the provided guidance and extensive feedback given by my supervisor Dr. K.J. Wolstencroft and the cooperation with fellow student Aster de Boer.

References

[Bri98] J P Brion. Neurofibrillary tangles and alzheimer's disease. *Eur. Neurol.*, 40(3):130–140, October 1998.

[BYBT10] Lynn M. Bekris, Chang En Yu, Thomas D. Bird, and Debby W. Tsuang. Review article: Genetics of alzheimer disease, 12 2010.

[CQ09] Kondethimmanahalli Chandramouli and Pei-Yuan Qian. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum. Genomics Proteomics*, 2009(1), December 2009.

[DA11] Vikas Dhikav and Kuljeet Anand. Potential predictors of hippocampal atrophy in alzheimer's disease. *Drugs Aging*, 28(1):1–11, January 2011.

[DK14] Doulaye Dembélé and Philippe Kastner. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics*, 15(1):14, January 2014.

[EDL02] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, January 2002.

[FSAL21] Qurat Ul Ain Farooq, Zeeshan Shaukat, Sara Aiman, and Chun-Hua Li. Protein-protein interactions: Methods, databases, and applications in virus-host study. *World J. Virol.*, 10(6):288–300, November 2021.

[Han21] Kristina Hanspers. Alzheimer's disease. https://www.wikipathways.org/instance/WP5124, June 2021.

[HCH+19] Yu-Ting Hu, Xin-Lu Chen, Shu-Han Huang, Qiong-Bin Zhu, Si-Yang Yu, Yi Shen, Arja Sluiter, Joost Verhaagen, Juan Zhao, Dick Swaab, and Ai-Min Bao. Early growth response-1 regulates acetylcholinesterase and its relation with the course of alzheimer's disease. *Brain Pathol.*, 29(4):502–512, July 2019.

[HLL+22] Luyan He, Xiaoman Liu, Hualian Li, Ruifang Dong, Ruobing Liang, and Ruoxi Wang. Polyrhachis vicina roger alleviates memory impairment in a rat model of alzheimer's disease through the egr1/bace1/app axis. *ACS Chemical Neuroscience*, 13(13):1857–1867, 2022.

[HZ06] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS genetics*, 2(6):e88, 2006.

[Jia22] C J R Jiang. *Finding Consensus Knowledge in the Huntington's Disease Pathway.* 2022.

[KBH09] Jungsu Kim, Jacob M Basak, and David M Holtzman. The role of apolipoprotein E in alzheimer's disease. *Neuron*, 63(3):287–303, August 2009.

[Lat93] David S Latchman. Transcription factors: an overview. *International journal of experimental pathology*, 74(5):417, 1993.

[LK14] Yun R Li and Brendan J Keating. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome medicine*, 6:1–14, 2014.

[LXL+21] Ziyi Luo, Hao Xu, Liwei Liu, Tymish Y. Ohulchanskyy, and Junle Qu. Optical imaging of beta-amyloid plaques in alzheimer's disease. *Biosensors*, 11(8), 2021.

[Men12] Mario F Mendez. Early-onset alzheimer's disease: nonamnestic subtypes and type 2 AD. *Arch. Med. Res.*, 43(8):677–685, November 2012.

[MLX+22] Xiangmao Meng, Wenkai Li, Ju Xiang, Hayat Dino Bedru, Wenkang Wang, Fang-Xiang Wu, and Min Li. Temporal-spatial analysis of the essentiality of hub proteins in protein-protein interaction networks. *IEEE Transactions on Network Science and Engineering*, 9(5):3504–3514, 2022.

[MNP+19] Naim Al Mahi, Mehdi Fazel Najafabadi, Marcin Pilarczyk, Michal Kouril, and Mario Medvedovic. GREIN: An interactive web platform for re-analyzing GEO RNA-seq data. *Sci. Rep.*, 9(1):7580, May 2019.

[NDB+18] Raffaella Nativio, Greg Donahue, Amit Berson, Yemin Lan, Alexandre Amlie-Wolf, Ferit Tuzer, Jon B Toledo, Sager J Gosai, Brian D Gregory, Claudio Torres, John Q Trojanowski, Li-San Wang, F Brad Johnson, Nancy M Bonini, and Shelley L Berger. Dysregulation of the epigenetic landscape of normal aging in alzheimer's disease. *Nat. Neurosci.*, 21(4):497–505, April 2018.

[NLD+20] Raffaella Nativio, Yemin Lan, Greg Donahue, Simone Sidoli, Amit Berson, Ananth R Srinivasan, Oksana Shcherbakova, Alexandre Amlie-Wolf, Ji Nie, Xiaolong Cui, Chuan He, Li-San Wang, Benjamin A Garcia, John Q Trojanowski, Nancy M Bonini, and Shelley L Berger. An integrated multi-omics approach identifies epigenetic alterations associated with alzheimer's disease. *Nat. Genet.*, 52(10):1024–1035, October 2020.

[PBM+06] Christina Priller, Thomas Bauer, Gerda Mitteregger, Bjarne Krebs, Hans A Kretzschmar, and Jochen Herms. Synapse formation and function is modulated by the amyloid precursor protein. *J. Neurosci.*, 26(27):7212–7221, July 2006.

[PMK+05] Yudi Pawitan, Stefan Michiels, Serge Koscielny, Arief Gusnanto, and Alexander Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024, 04 2005.

[PP10] Chandra Sekhar Pedamallu and Janos Posfai. Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information. *Source Code Biol. Med.*, 5(1):8, August 2010.

[PPM+05] Nikunj S Patel, Daniel Paris, Venkatarajan Mathura, Amita N Quadros, Fiona C Crawford, and Michael J Mullan. Inflammatory cytokine levels correlate with amyloid load in transgenic mouse models of alzheimer's disease. *Journal of neuroinflammation*, 2:1–10, 2005.

[PR14] Clara Pizzuti and Simona E. Rombo. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 01 2014.

[PTL20] Chi Him Poon, Long Sum Rachel Tse, and Lee Wei Lim. Dna methylation in the pathology of alzheimer's disease: from gene to cognition. *Annals of the New York Academy of Sciences*, 1475(1):15–33, 2020.

[PWG+15] Martin James Prince, Anders Wimo, Maelenn Mari Guerchet, Gemma Claire Ali, Yu-Tzu Wu, and Matthew Prina. World alzheimer report 2015-the global impact of dementia: An analysis of prevalence, incidence, cost and trends. 2015.

[QLC+20] Xuemei Quan, Huo Liang, Ya Chen, Qixiong Qin, Yunfei Wei, and Zhijian Liang. Related network and differential expression analyses identify nuclear genes and pathways in the hippocampus of alzheimer disease. *Med. Sci. Monit.*, 26:e919311, January 2020.

[RMN+05] M. Rentzos, M. Michalopoulou, C. Nikolaou, C. Cambouri, A. Rombos, A. Dimitrakopoulos, and D. Vassilopoulos. The role of soluble intercellular adhesion molecules in neurodegenerative disorders. *Journal of the Neurological Sciences*, 228(2):129–135, 2005.

[RRM+99] Wanda F. Reynolds, Jennifer Rhees, Dominique Maciejewski, Toni Paladino, Hans Sieburg, Richard A. Maki, and Eliezer Masliah. Myeloperoxidase polymorphism is associated with gender specific risk for alzheimer's disease. *Experimental Neurology*, 155(1):31–41, 1999.

[RSKS13] V Srinivasa Rao, K Srinivas, GN Sunand Kumar, and GN Sujin. Hypothesis protein interaction network for alzheimer's disease using computational approach. 2013.

[SA14] Tuba Sevimoglu and Kazim Yalcin Arga. The role of protein interaction networks in systems biomedicine. *Comput. Struct. Biotechnol. J.*, 11(18):22–27, August 2014.

[SBŠŠ11] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, July 2011.

[Smi98] Mark A. Smith. Alzheimer disease. volume 42 of *International Review of Neurobiology*, pages 1–54. Academic Press, 1998.

[SWH+93] Warren J Strittmatter, Karl H Weisgraber, David Y Huang, Li-Ming Dong, Guy S Salvesen, Margaret Pericak-Vance, Donald Schmechel, Ann M Saunders, Dmitry Goldgaber, and Allen D Roses. Binding of human apolipoprotein e to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset alzheimer disease. *Proceedings of the National Academy of Sciences*, 90(17):8098–8102, 1993.

[TAK+19] Sneham Tiwari, Venkata Atluri, Ajeet Kaushik, Adriana Yndart, and Madhavan Nair. Alzheimer's disease: pathogenesis, diagnostics, and therapeutics. *Int. J. Nanomedicine*, 14:5541–5554, July 2019.

[TJWJ11] Natalie A Twine, Karolina Janitz, Marc R Wilkins, and Michal Janitz. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by alzheimer's disease. *PLoS One*, 6(1):e16266, January 2011.

[TRI+13] Daniah Trabzuni, Adaikalavan Ramasamy, Sabaena Imran, Robert Walker, Colin Smith, Michael E Weale, John Hardy, Mina Ryten, and North American Brain Expression Consortium. Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.*, 4(1), November 2013.

[uni] UniProt ID mapping. https://www.uniprot.org/id-mapping.

[WN12] Malin Wennström and Henrietta M Nielsen. Cell adhesion molecules in alzheimer's disease. *Degener. Neurol. Neuromuscul. Dis.*, 2:65–77, July 2012.

[YLCC20] Xinjian Yu, Siqi Lai, Hongjun Chen, and Ming Chen. Protein–protein interaction network with machine learning models and multiomics data reveal potential neurodegenerative disease-related proteins. *Human Molecular Genetics*, 29(8):1378–1387, 04 2020.

[ZDJZ21] Ping Zeng, Jing Dai, Siyi Jin, and Xiang Zhou. Aggregating multiple expression prediction models improves the power of transcriptome-wide association studies. *Human Molecular Genetics*, 30(10):939–951, 02 2021.

[ZPZ+13] Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, and Bairong Shen. Biomedical text mining and its applications in cancer research. *Journal of biomedical informatics*, 46(2):200–211, 2013.

[ZTW+15] Xi-Chen Zhu, Lan Tan, Hui-Fu Wang, Teng Jiang, Lei Cao, Chong Wang, Jun Wang, Chen-Chen Tan, Xiang-Fei Meng, and Jin-Tai Yu. Rate of early onset alzheimer's disease: a systematic review and meta-analysis. *Ann. Transl. Med.*, 3(3):38, March 2015.

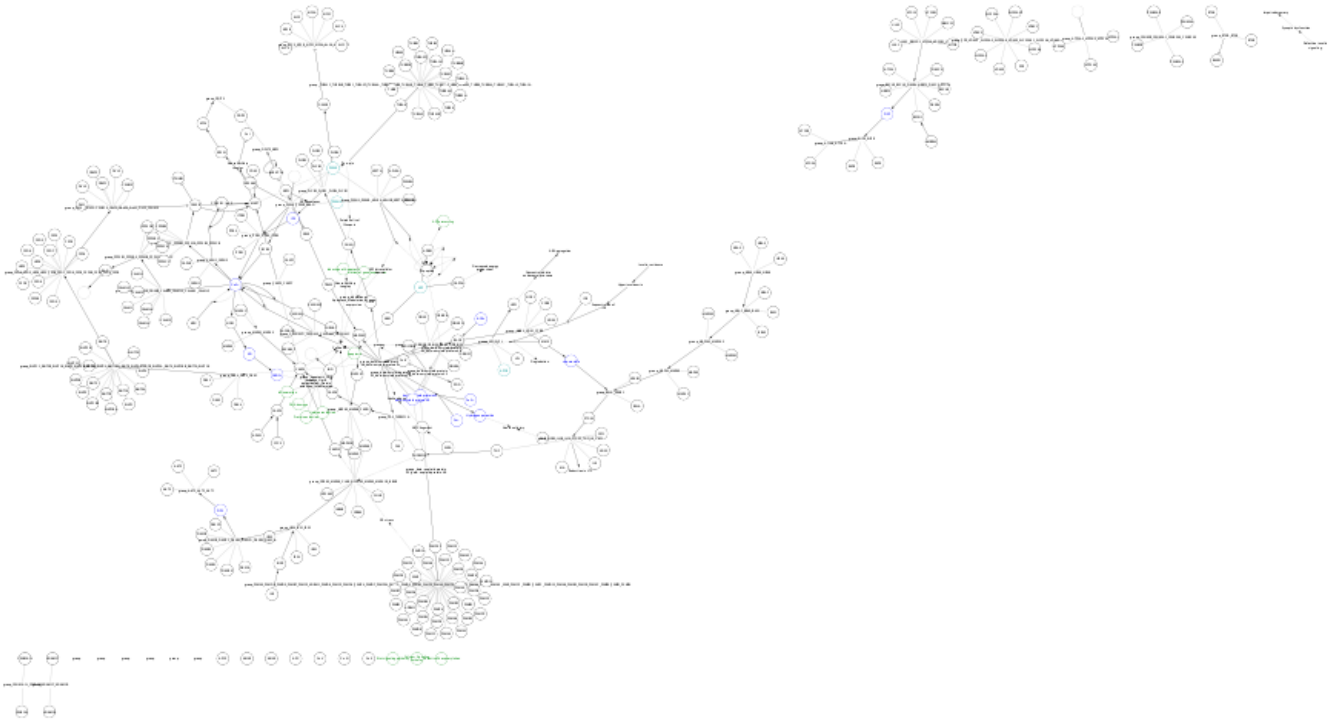


Figure 4: Visualisation of the WikiPathway PPIN without any modification.

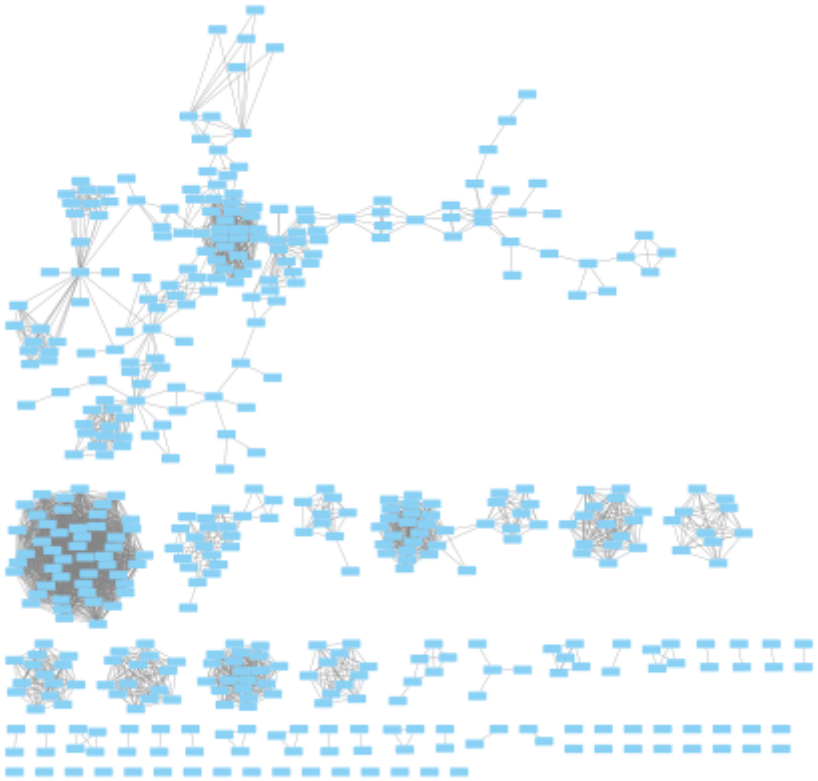
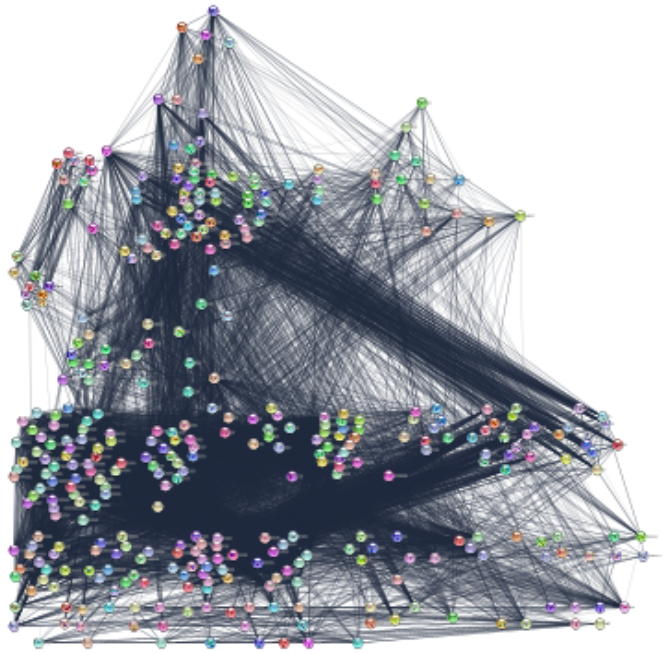
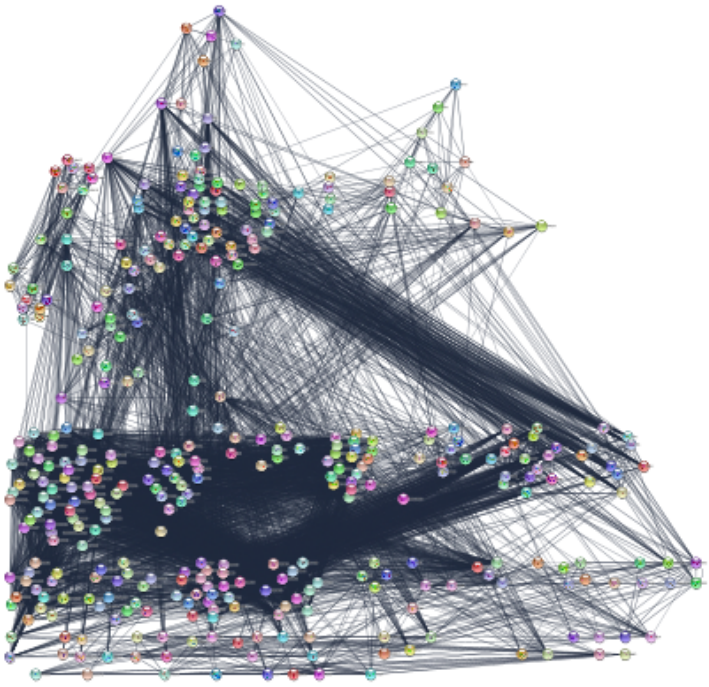


Figure 5: Visualisation of the KEGG PPIN, after modifications have been done by Aster de Boer.

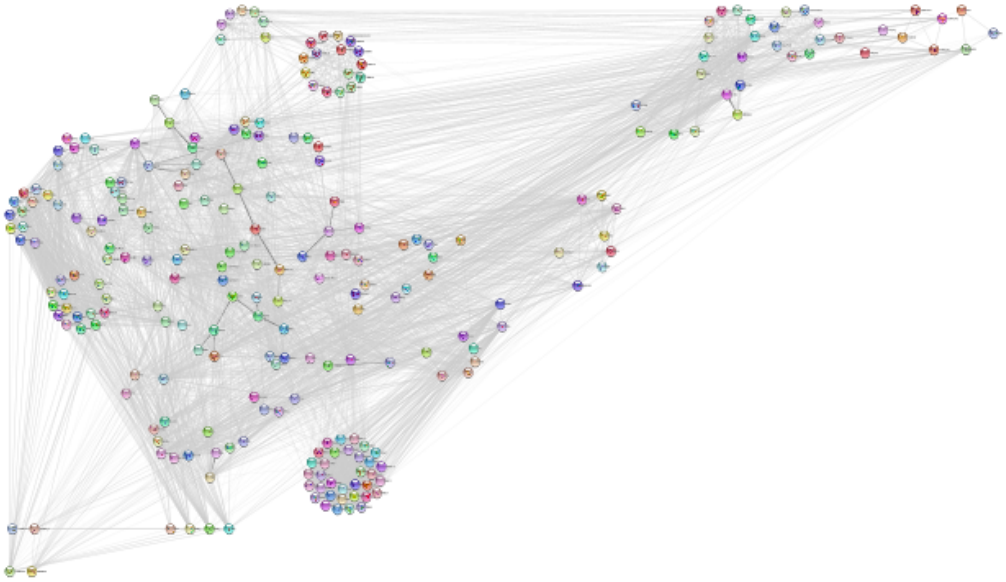


(a) Visualisation of the STRINGify KEGG network with a confidence cut-off score of 0.4.

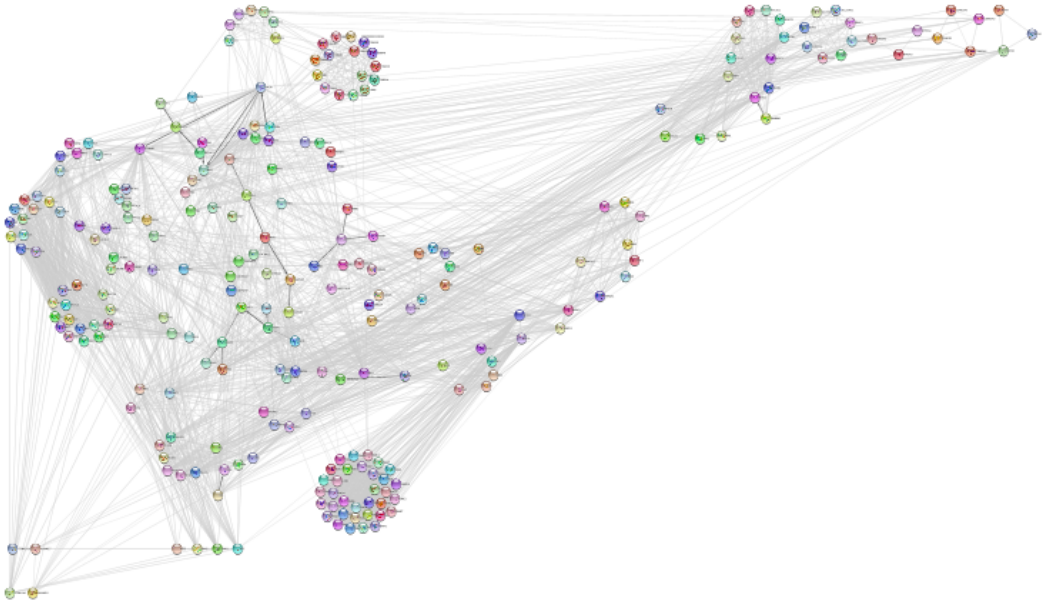


(b) Visualisation of the STRINGify KEGG network with a confidence cut-off score of 0.7.

Figure 6: Visualisation of the Stringify(0.4) and STRINGify(0.7) KEGG networks.

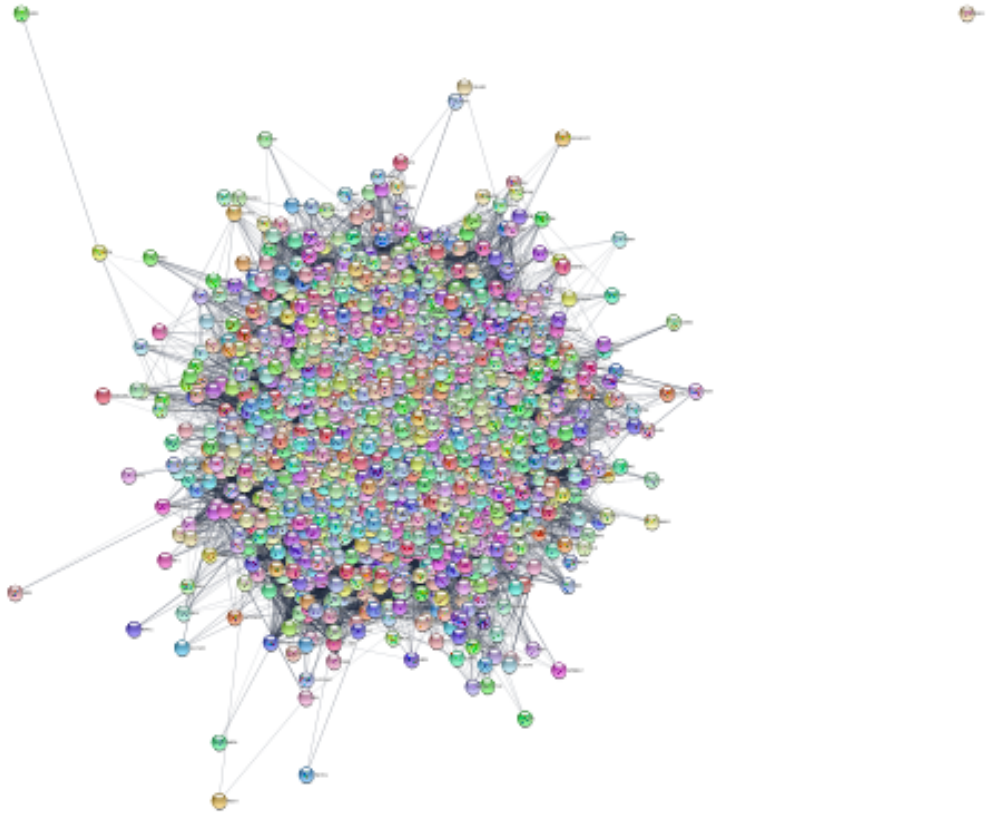


(a) Visualisation of the STRINGify WikiPathway network with a confidence cut-off score of 0.4.

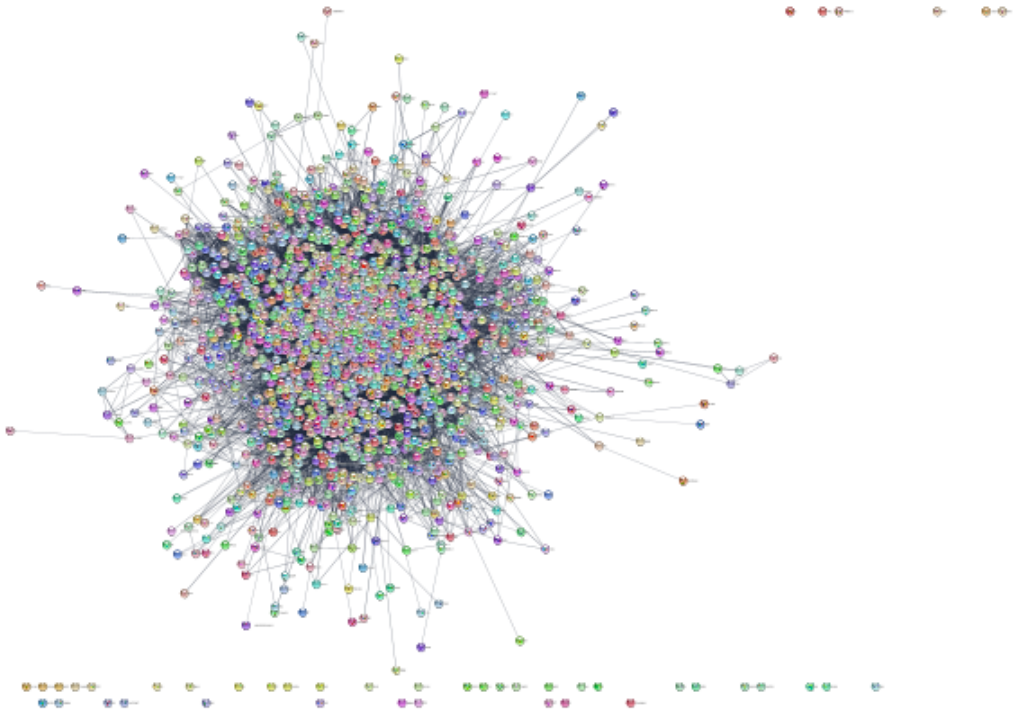


(b) Visualisation of the STRINGify WikiPathway network with a confidence cut-off score of 0.7.

Figure 7: Visualisation of the Stringify(0.4) and STRINGify(0.7) WikiPathway networks.

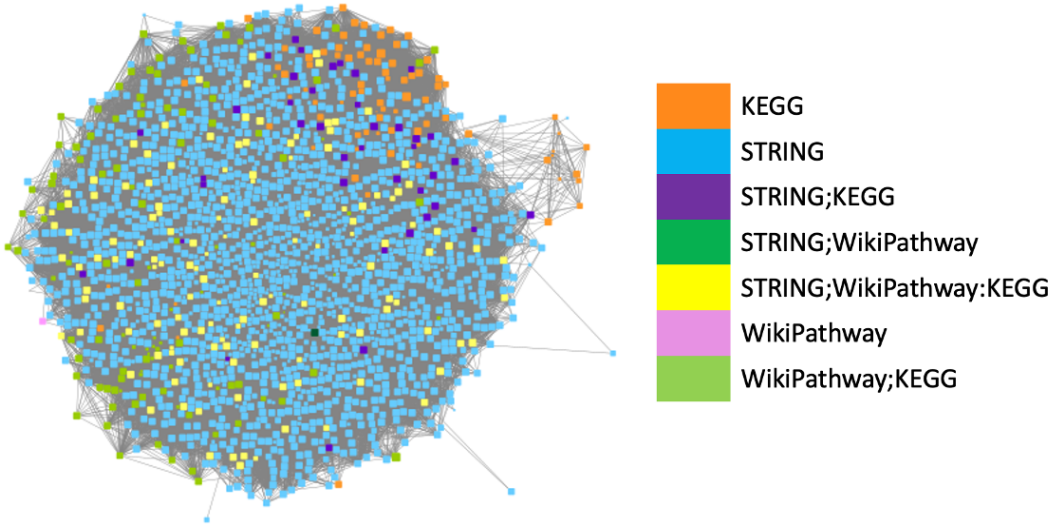


(a) Visualisation of the STRING network with a confidence cut-off score of 0.4.

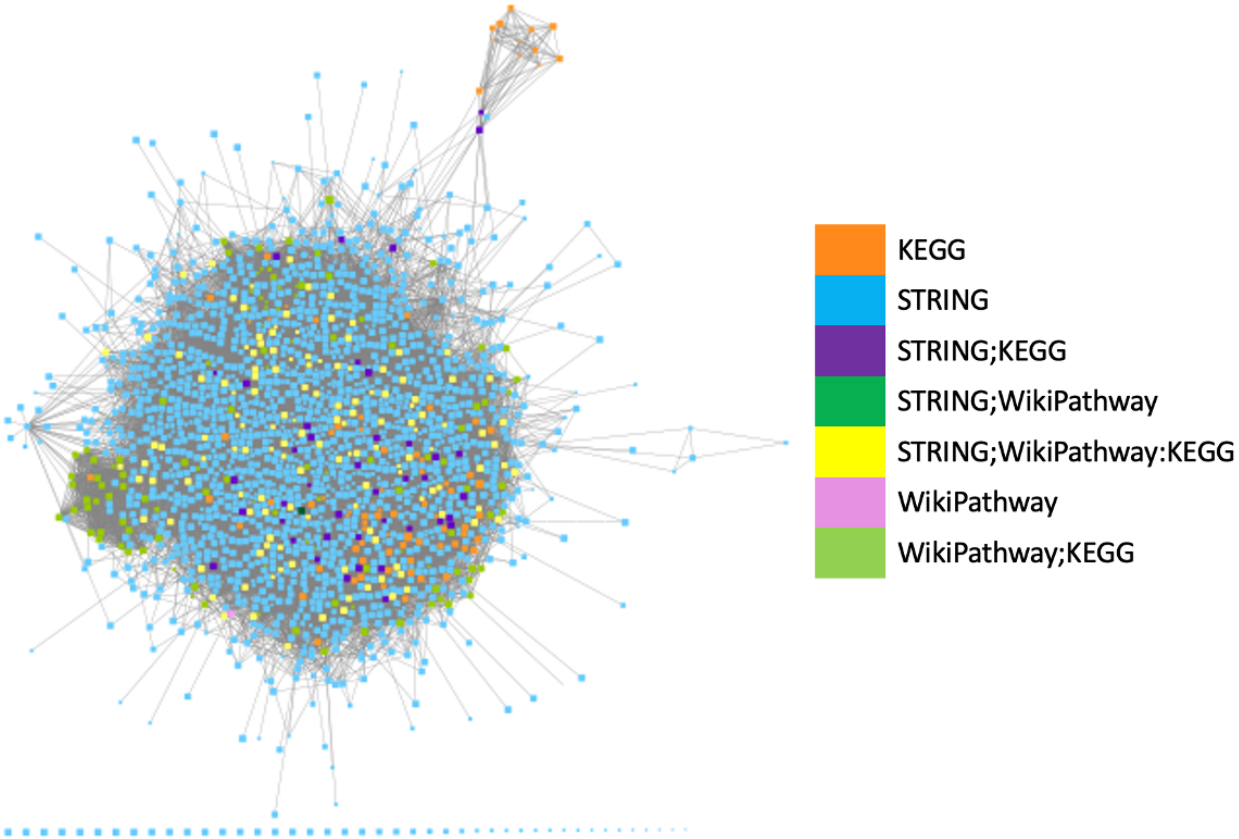


(b) Visualisation of the STRING network with a confidence cut-off score of 0.7.

Figure 8: Visualisation of the STRING(0.4) and STRING(0.7) networks.

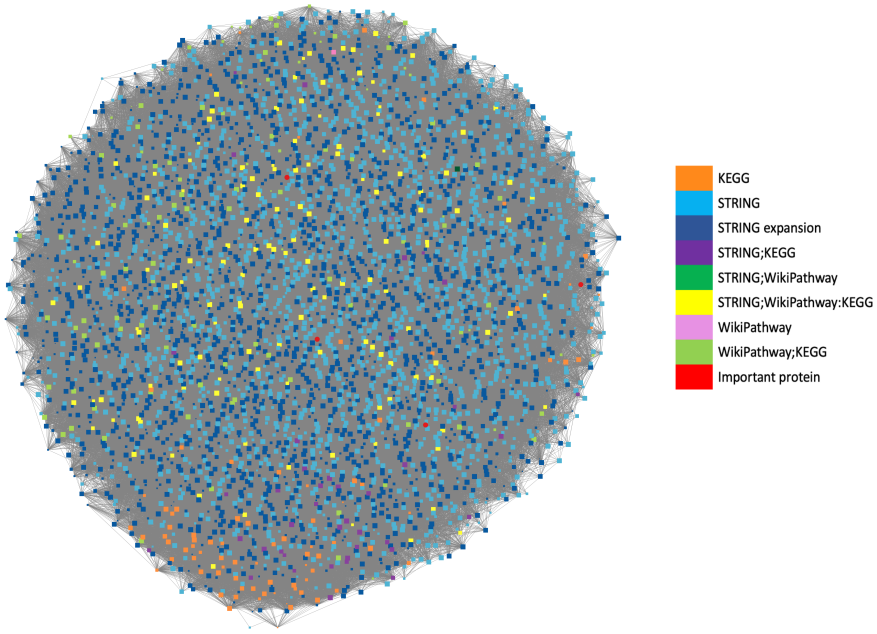


(a) Visualisation of the consensus(0.4) PPIN without STRING expansion.

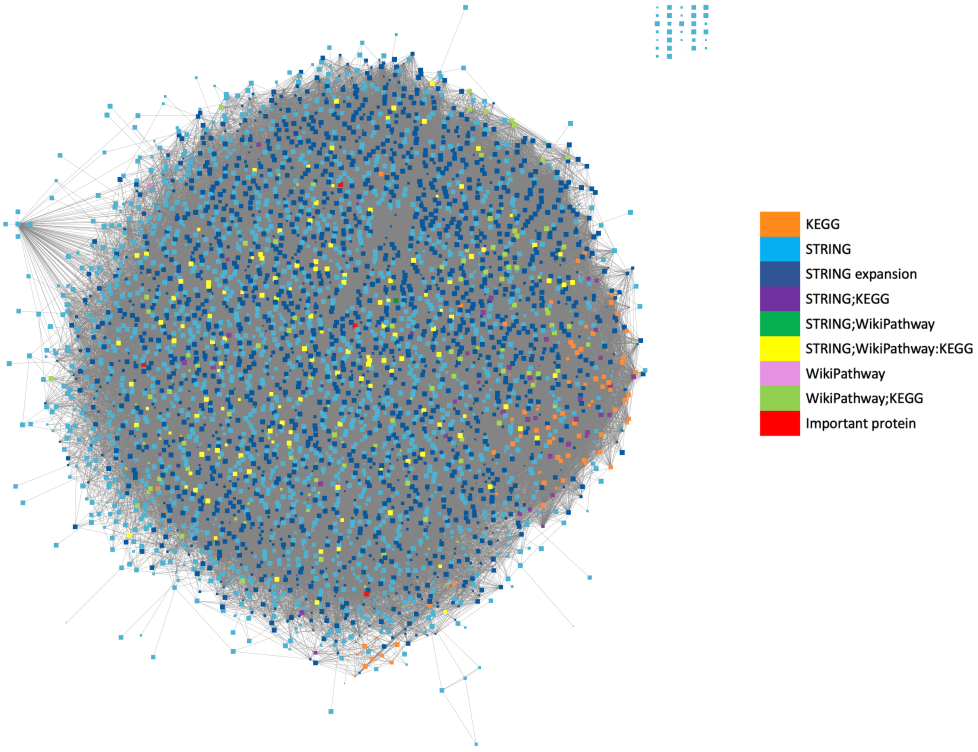


(b) Visualisation of the consensus(0.7) PPIN without STRING expansion.

Figure 9: Visualisation of consensus PPINs without STRING expansion.



(a) Expanded consensus 0.4 network.



(b) Expanded consensus 0.7 network.

Figure 10: Visualisation of the expanded consensus networks. The node size represents the protein's evidence score in the nervous system. The smaller the node, the less evidence there is that the node is present in the nervous system and vice versa.

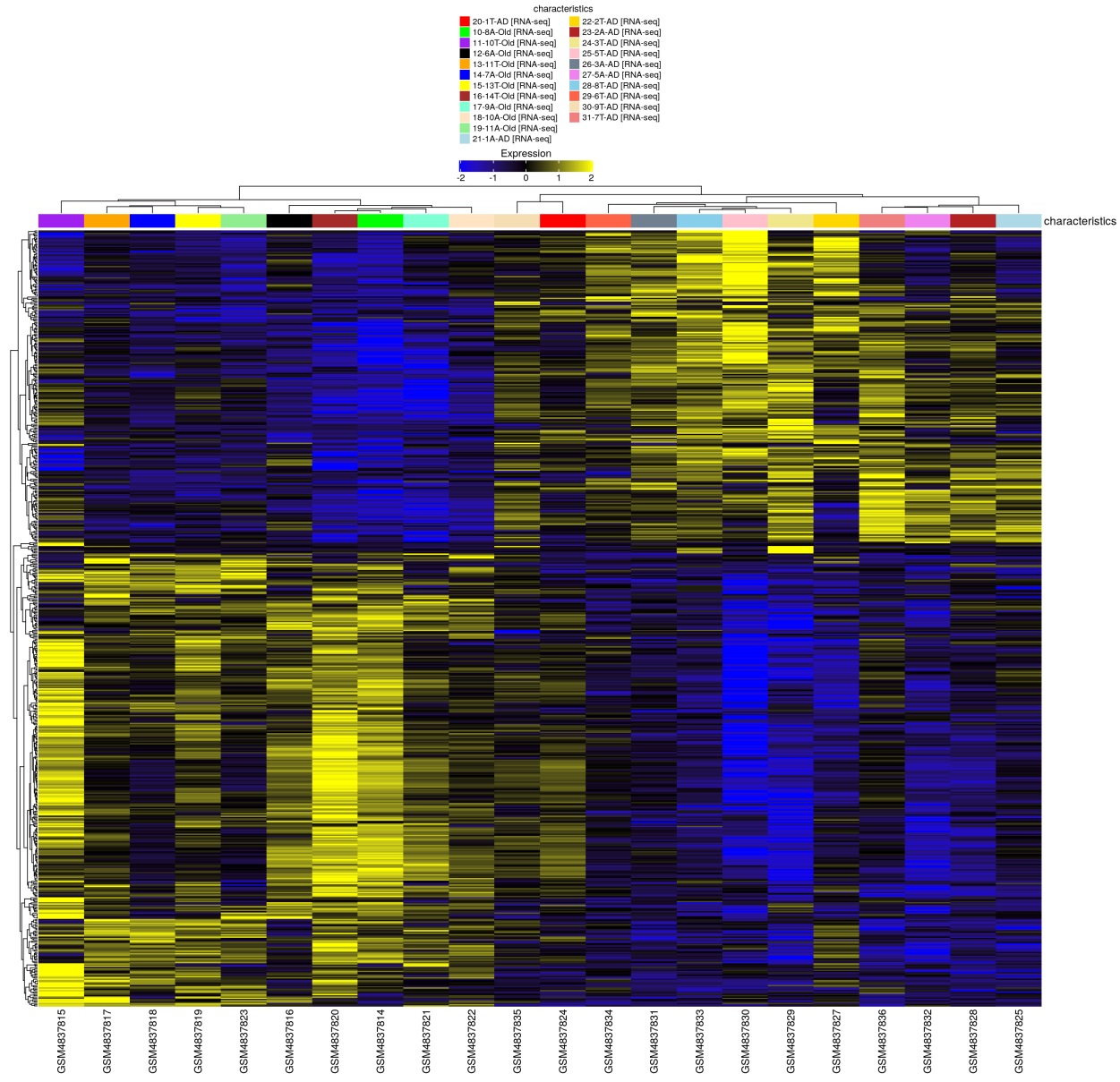


Figure 11: Heatmap of the top 610 differentially expressed genes in the GSE159699 dataset, obtained from GREIN.

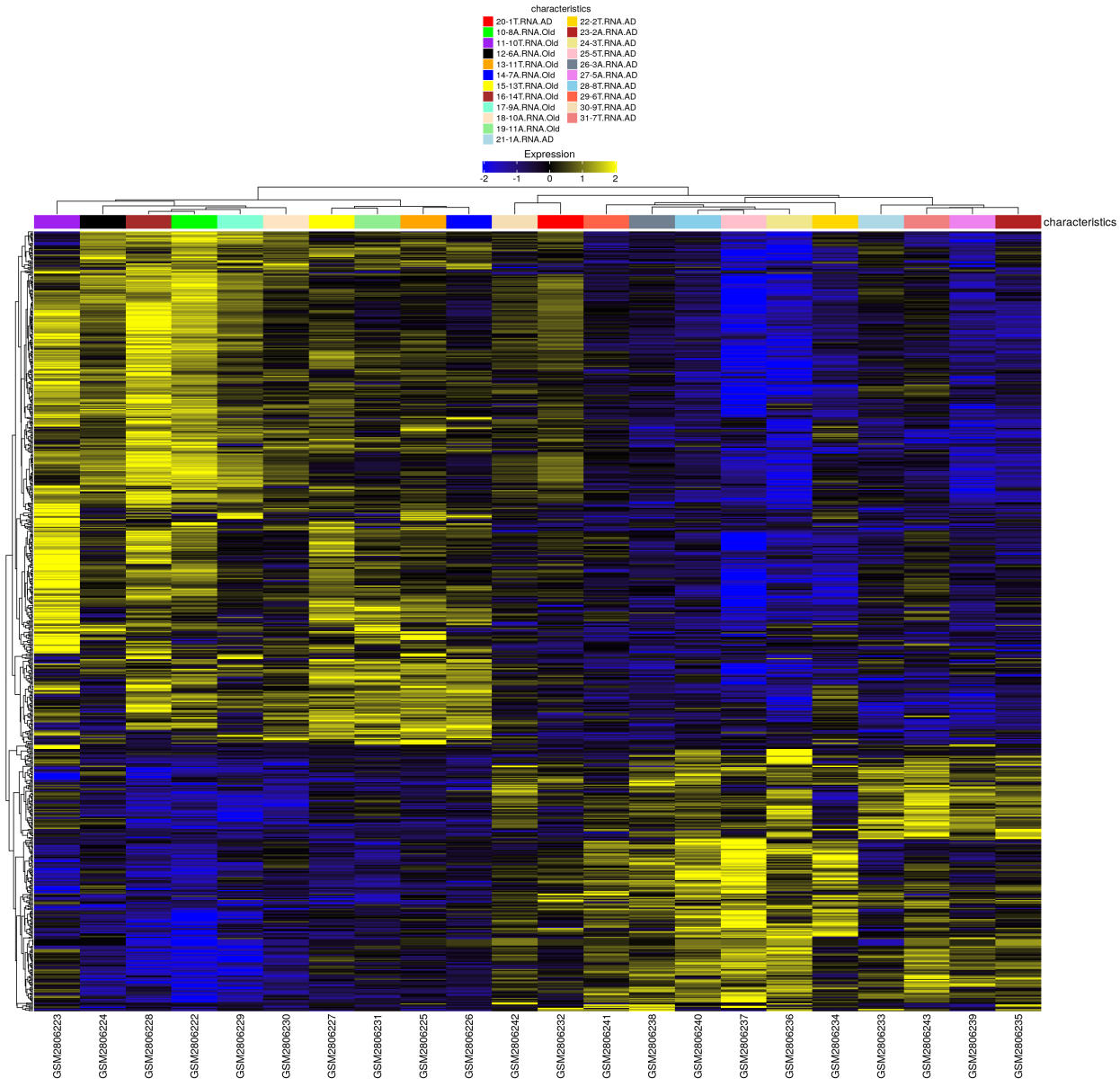
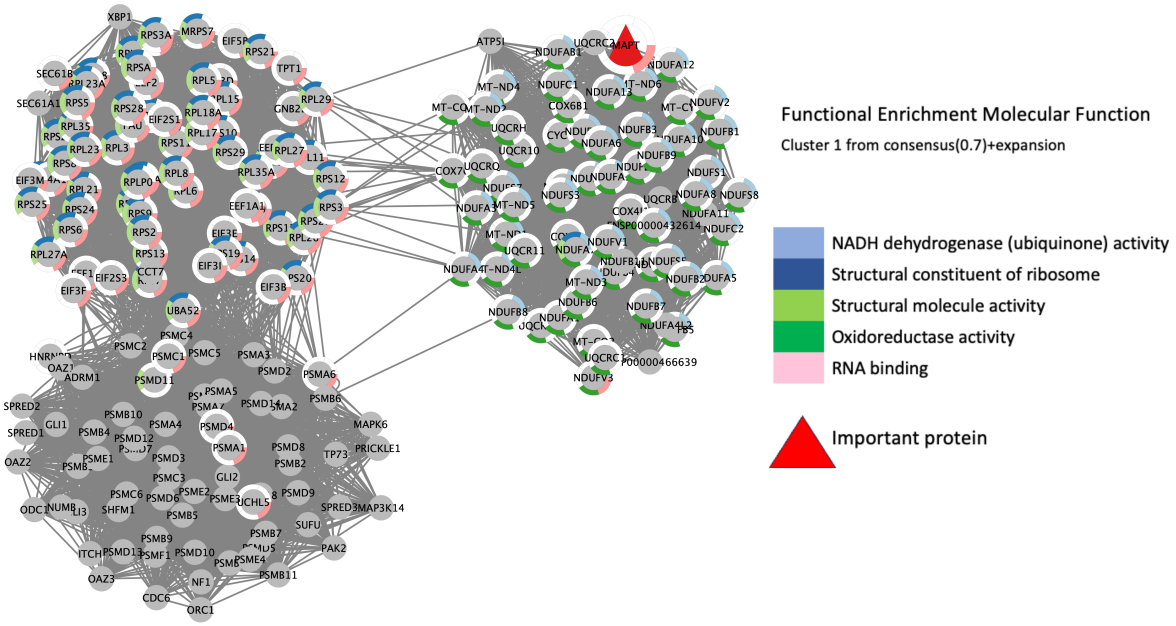
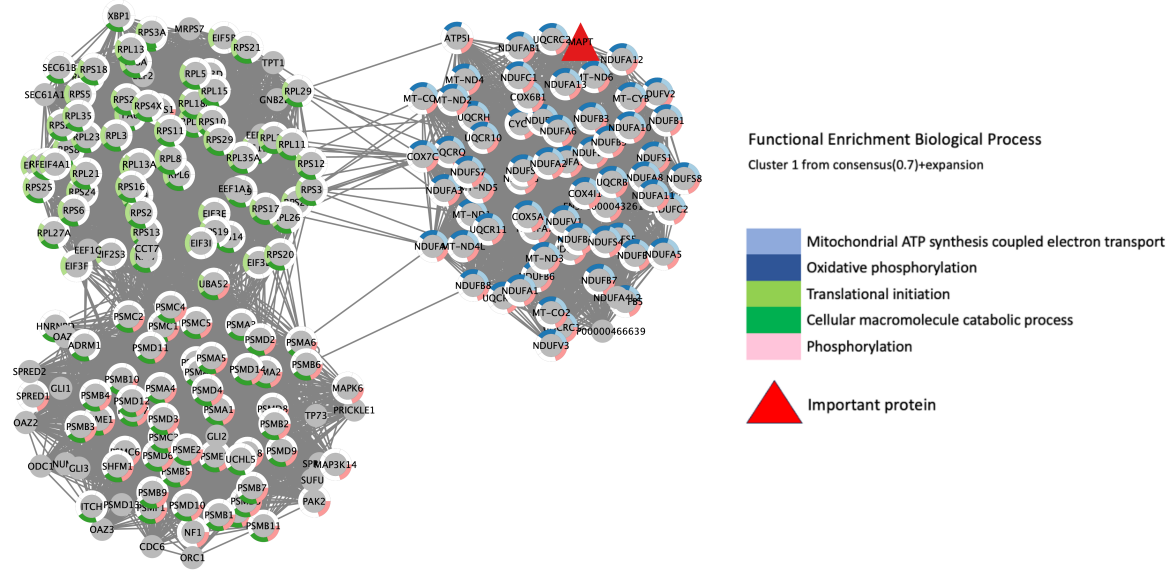


Figure 12: Heatmap of the top 702 differentially expressed genes in the GSE104704 dataset, obtained from GREIN.

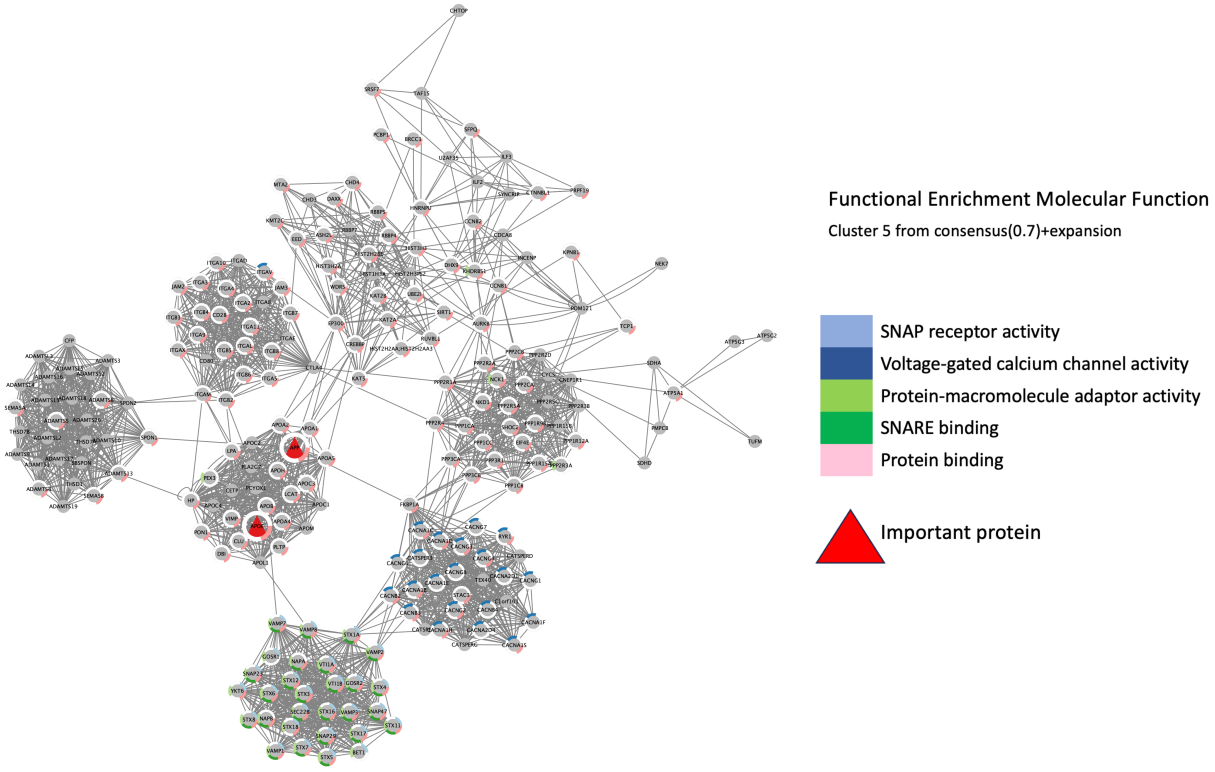


(a) Cluster 1 from the consensus(0.7)+expansion PPIN with their Molecular Function.

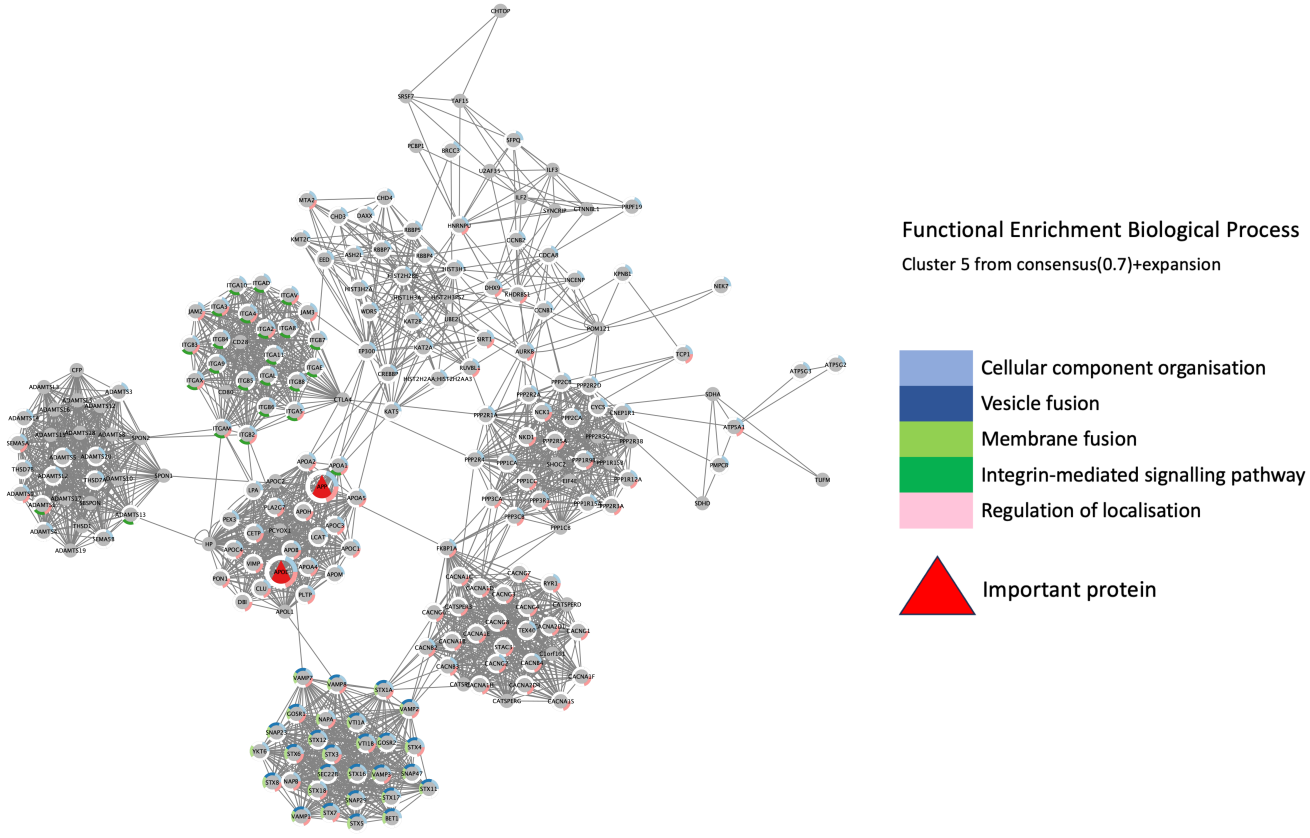


(b) Cluster 1 from the consensus(0.7)+expansion PPIN with their Biological process.

Figure 13: Functional enrichment analysis on cluster 1 from the consensus(0.7)+expansion.

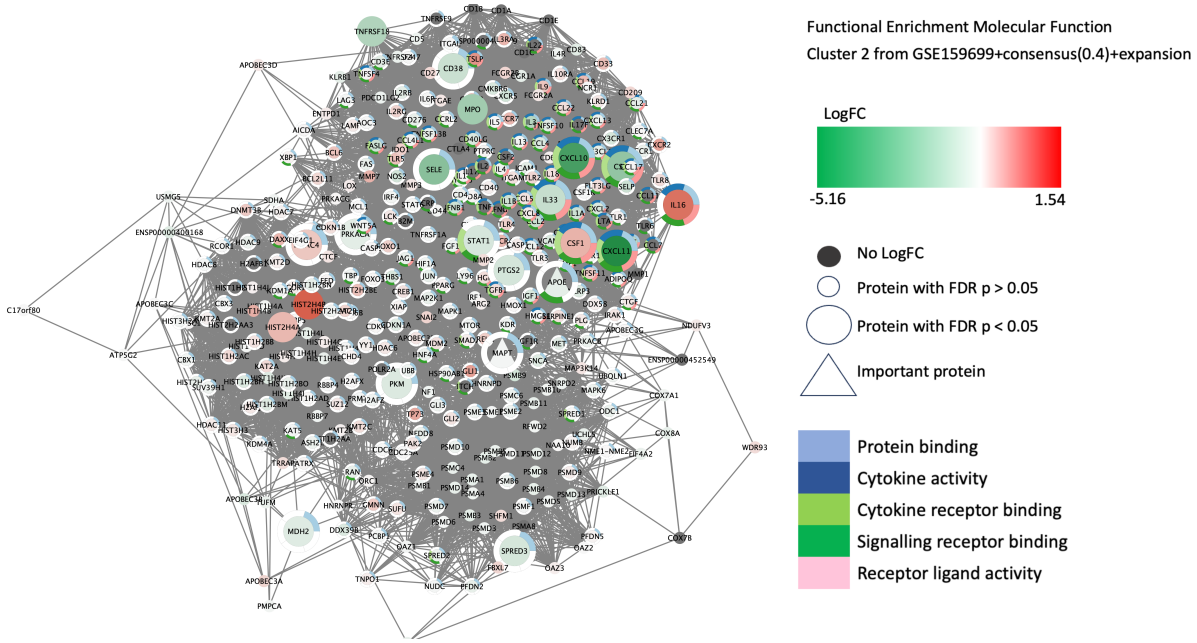


(a) Cluster 5 from the consensus(0.7)+expansion PPIN with their Molecular Function.

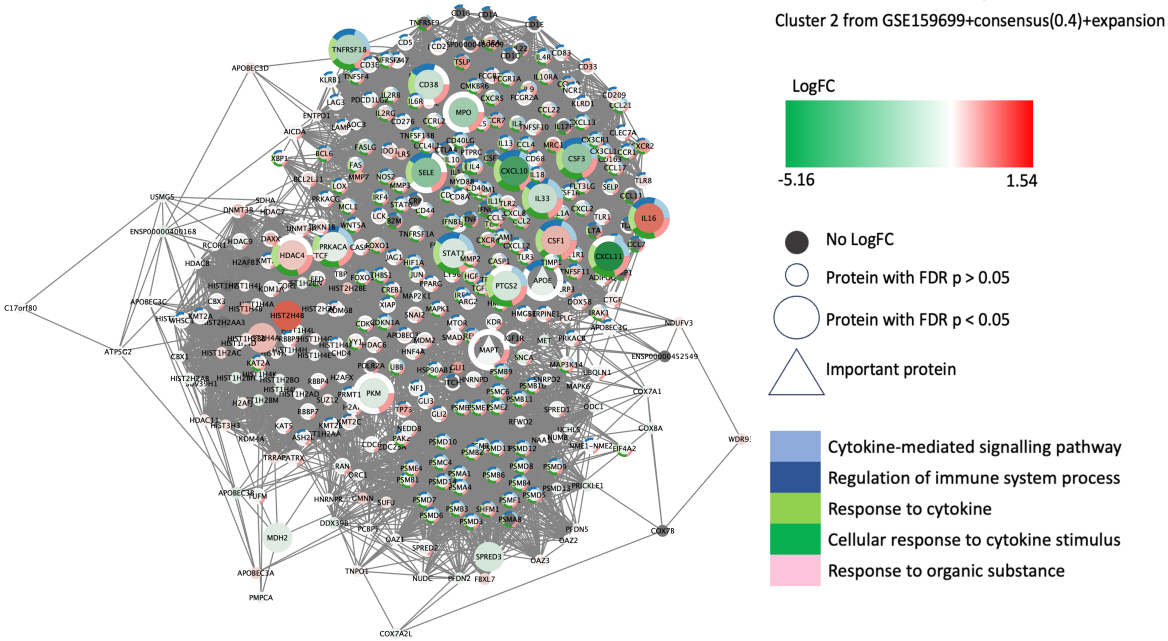


(b) Cluster 5 from the consensus(0.7)+expansion PPIN with their Biological process.

Figure 14: Functional enrichment analysis on cluster 5 from the consensus(0.7)+expansion.

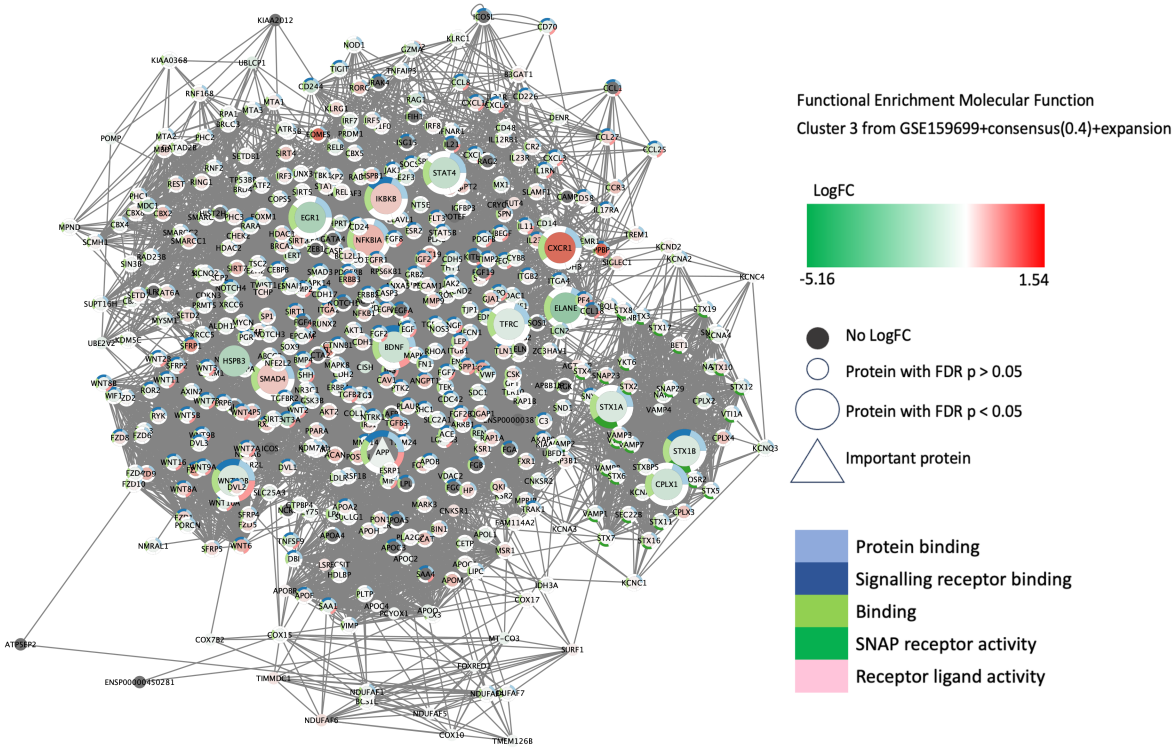


(a) GO MF on cluster 2 from the GSE159699+consensus(0.4)+expansion.

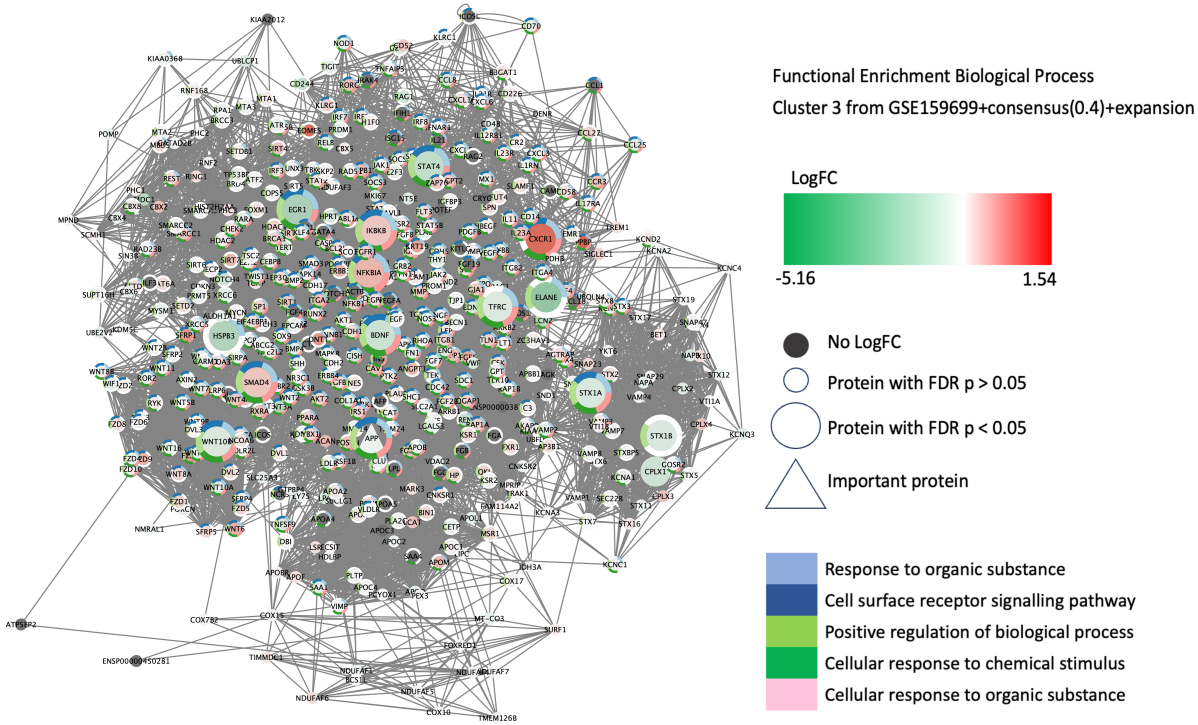


(b) GO BP on cluster 2 from the GSE159699+consensus(0.4)+expansion.

Figure 15: Enrichment analysis on cluster 2 from GSE159699+consensus(0.4)+expansion. The nodes are filled with a gradient, where green indicates downregulation and red indicates upregulation of the protein. The donut chart around the proteins represents the MFs or BPs in which this protein is involved.



(a) GO MF on cluster 3 from the GSE159699+consensus(0.4)+expansion.



(b) GO BP on cluster 3 from the GSE159699+consensus(0.4)+expansion.

Figure 16: Enrichment analysis on cluster 3 from GSE159699+consensus(0.4)+expansion. The nodes are filled with a gradient, where green indicates downregulation and red indicates upregulation of the protein. The donut chart around the proteins represents the MFs or BPs in which this protein is involved.

(a) GO MF on cluster 2.

(b) GO BP on cluster 2.

Figure 17: Tree map visualisation of the top 15 MFs and BPs present in cluster 2 from the consensus(0.4)+expansion PPIN. The default settings were used in Revigo.

(a) GO MF on cluster 3.

(b) GO BP on cluster 3.

Figure 18: Tree map visualisation of the top 15 MFs and BPs present in cluster 3 derived from the consensus(0.4)+expansion PPIN. The default settings in Revigo were applied.